uOttawa

L'Université canadienne
Canada's university

A Methodology for Reliable Data Mining on Health Administrative Data:

Case Studies on Pediatric Immune-Mediated Inflammatory Diseases in Ontario, Canada

by
Mohammad Hossein Tekieh

Thesis submitted to the University of Ottawa
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Digital Transformation and Innovation

Thesis directed by
Dr. Bijan Raahemi
Dr. Eric Benchimol

University of Ottawa
Ottawa, Ontario

April 2022

In the name of God,

the compassionate, the merciful

ABSTRACT

Over the past century, the prevalence of immune-mediated inflammatory diseases (IMIDs) has increased worldwide. It has been identified that exposures to environmental factors early in life are associated with increased risk of these diseases. However, hypothesis-driven analyses do not always identify all risk or protective factors, nor do they adequately explain interactions between variables on the risk of disease. Data mining has the capability of exploring the data without considering specific a priori hypotheses, instead providing possible hypotheses for further analysis. Though, data mining techniques are still not popular among epidemiologists as a trustworthy analytical tool to analyze population-based diseases due to inexplicability of some of the methods (e.g., neural networks), unfamiliarity with, or uncommon use of machine learning and data mining methods in real-world health care applications. At the same time, large amounts of routinely collected health data are amassed as a matter of operating electronic health systems. Routinely collected health data are not collected for research purposes; however, they are great sources of information for research as a secondary use of the data.

In this study, following the design science research methodology, we developed a methodology to reliably analyze health administrative data using data mining techniques to provide reproducible, reliable, and trustworthy findings. The reliable data mining methodology on health administrative data was designed in this study to address impartiality, validity, and sustainability concerns in five stages: Data Selection, Preprocessing, Modelling, Evaluation, and Feedback. As part of the main contributions, we developed two unique preprocessing guidelines as the key components of the designed methodology in order to standardize technical steps and address contextual sources of bias. While the proposed methodology is general in its design, to evaluate the designed methodology, we implemented it in several case

studies on the real health administrative data housed at ICES, Ontario, first to analyze children suffering with an IMID in Ontario, predict new cases, and, most importantly, generate new hypotheses. The first case study was extended to a second one to narrow focus from all IMIDs to asthma which formed the majority of the IMID cases. Eventually, a third case study was implemented focusing on inflammatory bowel disease (IBD) and systemic autoimmune rheumatic diseases (SARDs) to better compare the findings.

We applied both predictive and descriptive modelling techniques such as decision tree, neural network, logistic regression, and k-means clustering on the prepared datasets with more than 700K records and over 80 input variables. We built classification models with notable quality of performance (AUC of 68%), identified the significant factors associated to IMIDs, and extracted multifactorial rules causing protectiveness against or high risk of developing asthma, IBD, and SARDs. The factors that highly contributed to the extracted multifactorial rules were “general childhood infection”, “use of antibiotics”, “streptococcus pyogenes”, “respiratory infection”, “gastroenteritis”, “mother's prevalence of any IMID”, and “baby's sex”. The findings were evaluated and verified by health experts.

Most data mining studies which are applied to health data do not handle bias and confounding in their work. However, the systematic errors were identified, and their risks were assessed in these case studies due to following the designed reliable methodology. The results with high risk of bias were reported to disregard. Therefore, this process allowed us to apply data mining techniques to discover new multifactorial rules and identify the factors with the highest impact among the 128 factors observed in the past epidemiological studies, while preserving the trust of domain experts in the results.

ACKNOWLEDGEMENT

I would like to sincerely thank my supervisors, Dr. Bijan Raahemi and Dr. Eric Benchimol, for their extensive support and thoughtful directions through the peaks and valleys of this exciting journey. Also, I am grateful of my knowledgeable thesis advisory committee members, Dr. Doug Manuel and Dr. Liam Peyton, for their critical and constructive recommendations. Conducting this multidisciplinary research would not have been possible, without the patience and positive collaboration of experts from different fields of science.

I also would like to acknowledge that this research was financially supported by Mitacs Accelerate, Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery, and Children's Hospital of Eastern Ontario (CHEO) Research Institute.

Moreover, many thanks to my dearest parents for their continuous encouragements throughout my study and also in my life.

Last but not the least, I would like to thank my beloved wife, Farideh, from the bottom of my heart. Her patience, sacrifices, and caring in all aspects of our life, allowed me, inspired me, and motivated me not to give up and to continue.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

1 INTRODUCTION

Data mining integrates a set of multidisciplinary analytical methods from science and technology fields such as machine learning, statistics, databases, information retrieval, and artificial intelligence. It includes non-trivial techniques to discover meaningful and actionable hidden patterns in large amounts of data. Data mining methods are employed in many industries and sectors to build data-driven models based on their voluminous collected data. Although data mining is widely applied in the health care sector, there is often doubt regarding the validity of analyses, which may be mistrusted by clinicians, health administrators, policymakers, or public health specialists. In the era of digitization, data-driven systems are being switched from collecting hardcopy data to capturing digital data. This includes health systems. As a matter of course while operating electronic health systems, large amounts of routinely collected health data are amassed. Routinely collected health data are defined as data collected in the process of providing clinical care (such as data collected in electronic health records) or administering the health system (such as health administrative data). These data are not collected for research purposes; however, they are great sources of information for research as a secondary use of the data [1]. Health administrative data can be population-based. In Canada, these data are collected by provincial Ministries of Health in the process of administering the health system for all residents of the province. Population-based studies are conducted to answer research questions for a particular population, and the results should be valid and generalizable to the defined population [2]. Overall, it is believed there is a huge

potential of extracting useful information if there were a robust mechanism with which health administrative data could be reliably analyzed using data mining methods.

1.1 Motivation

Chronic illnesses are major threats to human health, affecting quality of life and leading to morbidity and mortality. They often require extensive and expensive investigation followed by treatment and can cause harm to people even in the pre-diagnostic phase. Over the past century, the prevalence of immune-mediated inflammatory diseases (IMIDs) has increased worldwide [3]. These diseases are triggered by a dysregulated immune response or auto-antibody formation, resulting in unexpected behavior of the immune system, which may damage different parts of body. The genetic origins of these diseases are well-described, but they are typically not monogenic. Many of these disorders are of higher prevalence in Westernized nations such as Canada, with low prevalence noted in people living in developing nations [4][5][6]. Recent studies of immigrants to Ontario have demonstrated increased risk with early life exposure to the Canadian environment amongst people from countries with low prevalence for these disorders [7][8]. In addition, exposures to other environmental factors early in life are associated with increased risk of these diseases. Some environmental factors found to be associated with inflammatory bowel disease (IBD), for example, include urban households, antibiotics, birth by Caesarean section, and air pollution [9]. However, hypothesis-driven analyses do not always identify all risk or protective factors, nor do they adequately explain interactions between variables or time-related effect modification of variables on the risk of disease.

Data mining is widely employed to discover the relationship between attributes hidden inside the data for decision making [10]. It has the capability of exploring the data without

considering specific a priori hypotheses, instead providing possible hypotheses for further analysis. Data mining is considered one of leading technologies with the potential to change the way we approach humanity's problems in various fields [11], including health care. The extracted information is useful for health specialists and will aid in making efficient decisions. Data mining techniques are divided into two main categories: descriptive (or unsupervised learning) and predictive (or supervised learning) [12][13]. Descriptive data mining is an exploratory analysis that attempts to measure the similarity of records and discover patterns and relationships. The most important techniques in descriptive data mining are clustering and association. On the other hand, predictive data mining techniques generate predictive rules as a model for classifying records based on a specific target (or label). Classification is the most widely used technique in predictive data mining.

Relying on its powerful analytical attributes, data mining can be a useful tool to assess disease risk and morbidity in large collections of routinely collected health data. In addition, routinely collected health data allow for the surveillance of rising chronic immune-mediated diseases such as IBD, asthma, and diabetes as well as traditional morbidity and mortality. Even though the reasons for the increased incidence of these chronic diseases are not yet clearly identified, the interaction between the environment, genetic predisposition, and host immune characteristics are thought to play a role. Therefore, assessing the interaction between these variables to predict risk using data mining techniques holds great promise for identification of the at-risk population and early intervention to prevent the disease or treat it promptly and efficiently. Having said that, data mining techniques are still not popular among epidemiologists as a trustworthy analytical tool to analyze population-based diseases due to uncertainty of the methods and unreliability of the results.

1.2 Context of the Research

To investigate the gap of employing data mining as a *technical field* in epidemiology (the science of studying disease patterns in human populations) as the *contextual field*, a multidisciplinary study is needed. The world is changing significantly due to electronic digital technology implementation in an expanding set of sectors and industries. The collection and availability of large amounts of data have contributed to the development of novel data analysis methods. This change transforms how business and society work. Therefore, new ideas and innovation are needed to provide new solutions that take advantage of electronic digital technology to solve more problems.

The Digital Transformation and Innovation program seems to be a perfect fit to define a research study addressing the concerns and potentials raised in the Motivation section. The program allows studying the interaction between various fields from the electronic technology, business, and society pillars. To do so, the researcher should investigate the characteristics of each focused field to better understand its details and subsequently design a new approach or solution to build a bridge between them so that each can take advantage of the others' benefits, while addressing the challenges in between.

1.2.1 Technical Aspect

Data mining and machine learning algorithms have advanced significantly in the academic world in the past few decades. In the recent years, they have also received extraordinary attentions from various sectors and industries in which their applications have triggered remarkable number of innovative initiatives.

In this study, we plan to investigate the details of an established data mining process and enrich it by adding additional components in order to provide valid findings specifically when analyzing health administrative data. The technical achievements and outcomes will enable data mining techniques to be used with more confidence in this particular context.

1.2.2 Business Aspect

Even in regions where health care systems are funded by governments, managing the financial resources in both the public and private health sectors is essential for sustaining any administrative system. This makes the context of this study — population health — noteworthy from a business perspective. For instance, Canada was expected to spend a quarter of a trillion dollars annually on health care by 2020 [14]. In 2011, more than 50% of the total government budgets in the provinces of Ontario and Quebec was consumed for provincial health spending, and this number is still increasing [14]. Huge budgets are allocated to health care with the majority used by a small portion of the population [15]. It is thought that understanding the patterns and predicting which patients will develop chronic diseases — which put a highly expensive burden on the health care system — could lead to prevention, early intervention, and reduction of health costs. In this thesis, we are particularly motivated to target children with immune-mediated diseases in Ontario, Canada, to conduct an experiment using the proposed solution. Recent studies show the rates of inflammatory bowel disease (IBD), multiple sclerosis (MS), asthma, and type 1 diabetes mellitus (T1DM) have increased significantly in Canada [16], thereby increasing the annual health cost on the shoulder of public and private health sectors.

1.2.3 Social Aspect

In addition to the business aspect of the focused context — population health — it is understandable how any topic related to health care could impact the quality of life of a society's residents as well. Investigating the relationship between improvement of a population's health status and social impacts is out of the scope of this study. However, we believe that achieving practical outcomes by building a bridge between a technical ability (data mining) and the focused context (population health) will elevate the quality of a society's health services. In Canada, the enormous health care expenditure has not consistently been accompanied by improvements in quality of health services and delivery. Among 11 developed countries surveyed, Canada has the worst access times to a doctor or a nurse when in need of care, the most extreme delays for specialist appointments, and the highest use of emergency department services. In addition, in terms of medical error incident rate, Canada was ranked 7th among 15 peer countries [14].

1.3 Objectives

In this study, we aim to design a methodology to reliably analyze health administrative data using data mining techniques. There are contextual considerations included in many epidemiology studies in order to provide valid and reliable results. Normally, these considerations are not taken into account in studies that apply data mining to health data. Therefore, the main objective is to design a reliable methodology that provides valid and trustworthy findings — from the perspectives of both data miners and epidemiologists — by mining health administrative data.

The second objective is to conduct experiments that will apply the designed methodology to real-world health issues using data from children diagnosed with IMIDs. The root causes of

these disorders are unknown. Our study population will consist of children because they have been exposed to fewer environmental factors, which will reduce the complexity of the variables' interrelationships and increase the chance of generating new hypotheses concerning links between environmental factors and disease risk.

1.4 Research Questions

Based on the objectives of this study, we defined the following research questions:

1. What are the missing parts in a typical data mining study that are accounted for in many epidemiological studies in order to provide trustworthy results?
2. Can we design a reliable data mining methodology to produce valid findings when analyzing health administrative data while addressing the related epidemiological considerations?
3. How can this methodology be applied to health administrative data to generate new hypotheses on the association between IMIDs and early life and environmental factors?
4. What are the significant early life and environmental factors that are associated with increased risk of IMIDs using data mining techniques?

1.5 Methodology

To answer the above research questions, we applied a data science research methodology to build a reliable data mining methodology as the outcome of this study. As we were developing an innovative, purposeful, and viable artifact that other professionals can use to design solutions for their problems, the design science research methodology in information systems was selected. In addition, we evaluated this artifact (i.e., the reliable data mining methodology

to analyze health administrative data) by running it on real-world health data to analyze a rising population-based health disorder as a case study. Finally, the results of the research were presented to technology-related and health-related audiences. This research study complies with the seven guidelines that Hevner et al. [17] count for a design science research, including design as an artifact, problem relevance, design evaluation, research contributions, research rigor, design as a search process, and communication in research.

It is worth mentioning that the studies conducted as a matter of implementing the proposed methodology — such as the case studies we conducted — are mainly quantitative research studies. These studies attempt to model phenomena by employing mathematical theories. More specifically, analysts using the proposed methodology will build valid models by applying data mining techniques to health administrative data to investigate a health-related issue.

1.6 Thesis Contributions

The major contributions of this thesis are:

1. Clarifying the terminology. Both data mining and epidemiology are multidisciplinary fields with some common ground, such as in statistics. The terms used to express a concept can be different in different disciplines as they have different perspectives. A problem arises when terms are used interchangeably in a field that aggregates those disciplines, and there are many terms in data mining and epidemiology that refer to the same concepts. Especially as we try to provide or suggest a specific perspective in conducting the studies, we determine the terms that refer to the same concept and distinguish the differences.

2. Designing a new data mining methodology. Based on our initial investigation, there were three key features that were missing in order to make a data mining methodology reliable from an epidemiological perspective: the method must be unbiased, valid, and sustainable. We do not want to reinvent the wheel; therefore, we start with CRISP-DM (CRoss Industry Standard Process for Data Mining), which is a well-known data mining process. We recategorize its stages and enhance the process by adding new components specifically designed for our focused context. The proposed solution is validated and improved by running case study experiments. This new methodology is designed to run data mining reliably on health administrative data.
3. Forming a technical preprocessing guideline. There are many different preprocessing actions suggested in data mining resources. However, we wanted the sequence and actions to be noticeably clear, customized for analyzing health administrative data, and applicable to all the selected descriptive and predictive techniques. The steps were aggregated carefully in the format of a guideline that we refer to as the Technical Preprocessing Guideline (TPG).
4. Creating a contextual preprocessing guideline. One of the main contributions of this thesis is this preprocessing guideline, and it was revised multiple times to become as practical and robust as possible. The main epidemiological considerations that made most earlier data mining works untrusted are discussed and resolved step-by-step in a unique preprocessing approach entitled Contextual Preprocessing Guideline (CPG).
5. Building predictive models for IMIDs. The designed methodology was implemented on real-world data to analyze pediatric patients suffering with an IMID in Ontario, predict new cases, and, most importantly, generate new hypotheses by evaluating the most important variables. The case study was extended to a second one to narrow focus

from all immune diseases to a prevalent one (asthma) which formed the majority of our immune disease cases. Eventually, a third case study was implemented focusing on IBD and systemic autoimmune rheumatic diseases (SARDs) which includes juvenile idiopathic arthritis (JIA) patients to better compare the findings. Therefore, all stages crafted into the new methodology were completely implemented: retrieving risk and protective factors, selecting databases, creating the dataset, technical and contextual preprocessing of data — which requires the most effort in the process — building various predictive models, evaluating the experimental models, and validating the results.

1.7 Publications

To date, the following book chapter and conference paper have been published from this study:

- Tekieh, M. H., Raahemi, B., Benchimol, E. I., (2017) "**Leveraging Applications of Data Mining in Healthcare Using Big Data Analytics: An Overview**", Chapter 15, Handbook of Research on Data Science for Effective Healthcare Practice and Administration, Noughabi, E. Raahemi, B., Albadvi, A., Far, B. (Edt.) IGI Global, Pages 345-359.
- Tekieh, M. H., Raahemi, B., (2015) "**Importance of Data Mining in Healthcare: A Survey**" Proceedings, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM2015), Paris, France, pp. 1057-1062, August.

1.8 Thesis Structure

In the first chapter of this thesis, in addition to the health problem that motivated us to conduct this research study, the research context, objectives, questions, methodology, and main contributions are briefly described. In Chapter 2, essential background information and a literature review are presented to better understand the grounds of this research study. The research methodology employed in this study is presented in Chapter 3, while the designed data mining methodology is elaborated in Chapter 4 as the main outcome and contribution of this work. In Chapter 5, the technical and preprocessing guidelines are presented as important contributions. Chapters 6 to 8 report the case studies wherein the designed data mining methodology comes into real-world action to build actual predictive models using Ontario pediatrics data to analyze IMIDs. Finally, Chapter 9 concludes with the summary of work along with potential future works.

CHAPTER TWO

2 BACKGROUND

Epidemiology is the study of population health. In traditional epidemiology, statistical methods are employed to run analyses on collected and prepared data. Typically, a population sample is selected or collected to test a predefined hypothesis using statistical analysis. In cases where the entire population's data are available, feasibility is a huge concern even though statistical methods can mathematically analyze them. With large datasets, a correspondingly large amount of computing power is required for statistical analysis, and achieving results in timely manner is still under serious doubt due to the nature of the methods that precisely calculate similarities and dissimilarities. On the other hand, data mining techniques use artificial intelligence and heuristic functions which are based on fuzzy calculations. These methods perform much better on large volumes of data and include all available information in the analysis, rather than a sample of the data. In this study, we aim to employ data mining techniques to analyze health data. These techniques are hypothesis-generating and can also be augmented using big data analytics. However, big data platforms might not always be necessary, depending on the type and amount of data, which we discuss this further in section 2.4.4.

As discussed, various fields of science are involved and integrated in this study. Therefore, it is essential to define, clarify, and briefly discuss the terms and concepts used in this study. In addition, we review past works related to the topic. We start with better understanding our main material — health data. Then, we continue with some introduction to and discussion of

related investigations in epidemiological studies and data mining. Finally, we discuss challenges and suggestions related to the terminology of these concepts.

Note that sections **2.1.2** (brief discussion on different types of health data), **2.4.1** (brief discussion on applications of data mining to health care), and **2.4.4** (comprehensive discussion on the potential role of big data analytics in boosting applications of data mining to health care) have been published before in our book chapter [18] with DOI of 10.4018/978-1-5225-2515-8.ch015. Also, sections **2.3.1** (brief introduction of data mining techniques), **2.3.3** (advantages of data mining over traditional statistics), **2.4.2** (brief discussion on the necessity of data mining in health care), and **2.4.3** (brief discussion on the challenges of using data mining in health care) have been published before in one of our papers [19] with DOI of 10.1145/2808797.2809367. Both publishers have provided authors the right to reuse the published texts in their dissertation.

2.1 Health Data

The more health care systems switch to electronic digital technologies, the more health data are generated. Health care systems are broad and include many different aspects such as laboratory tests, monitoring devices, hospital discharges, bedside clinical assessment, billing, insurance, staff administration, and online consultations. Analysts usually tend to think of data in a very standardized and structural way. We clearly observe this mindset with studies using any sort of health data to validate new analytical algorithms and easily claiming the results as new findings. However, we believe that health data are not as easy to work with as initially thought. In this section, we introduce some challenging characteristics of health data to provide insights on this matter. In addition, we discuss the various types of health data in an organized fashion to help with planning for them wisely and analyzing each type properly.

2.1.1 Challenging Characteristics

LeSueur [20] has discussed some characteristics of health data that makes them unique and difficult to measure. Figure 2.1 summarizes the issues that makes analyzing health care data difficult. Those challenges, and how they make health data so uniquely complex, are discussed briefly below.

Health care is hugely different from industrial businesses where definitions and rules are stable for a long period. Dealing with all these inconsistencies and complexities is not easy [20].

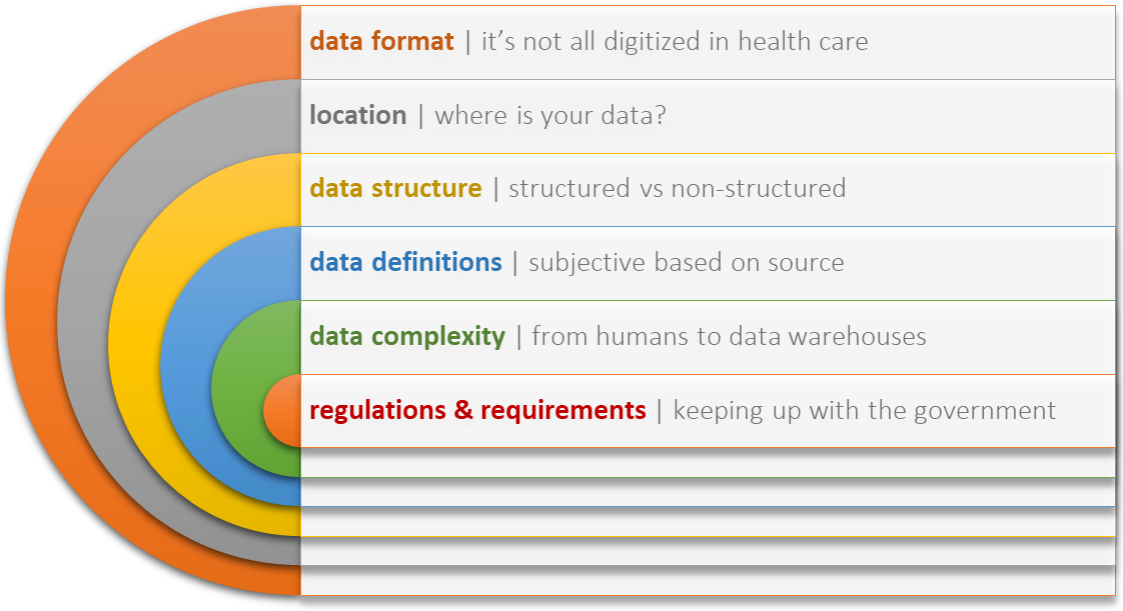

Figure 2.1: Why health care data are difficult [20]

2.1.1.1 Data Format

The collected data can be stored in various formats, depending on the source. The radiology lab records the medical condition in an image. The claim record captures the same condition in ICD codes. Different organizations record diagnoses using different coding systems, such as ICD-9, ICD-10, or even custom versions. Clinical data can be collected electronically via

an electronic medical record (EMR) system, while older medical records may still be in paper copies.

2.1.1.2 Location

Health data are housed in multiple sources. Even in one health care provider organization, such as a hospital, health data can come from different places — such as HR, clinics, the ER, and labs — using totally different database structures. The same issue could be duplicated in multiple sources with different formats, for example, an image in the radiology lab and an ICD-10 code in the claim records. Aggregating them into a data warehouse could make the data both accessible and usable.

2.1.1.3 Data Structure

EMR systems were implemented to provide facilities in instant access to previous records, preferably in a standardized and structured manner. However, the captured data are far from being consistent as there are still opportunities to enter unstructured data in the form of free-form text, for example, when providers are trained to capture the details in whatever way that is most convenient for them in order to speed up the process and record the required information. On the other hand, EMR systems sometimes provide different features referring to the same clinical concept without much standardization or even do not offer enough flexibility to capture precise information; for example, there may be limited options to select in a field. Therefore, this results in the systems failing to achieve integrated, structured, and standard data collection. Improvements in EMR software applications and more training in their timely use would slowly reduce this challenge.

2.1.1.4 Data Definitions

There might not always be consensus about the definition of a cohort or treatment between different groups of clinicians. Each follows separate criteria to diagnose diseases; therefore, collected health data from different sources can face inconsistency in case identification leading to misclassification bias. This issue can be handled to some extent by using validated regional algorithms. However, due to the nature of the medical science wherein new discoveries are frequently made, the agreed-upon knowledge continually evolves. Therefore, definitions are variable, making it challenging to compare data in a timely fashion. If captured data are recorded in a raw format, this could make it possible to identify cases, both current and historical, using the latest criteria and to form a more consistent cohort. Nevertheless, having access to raw diagnosis codes is still not ideal since different clinical approaches could result in different diagnosis codes, in addition to the risks of miscoding – entering the incorrect diagnosis code.

2.1.1.5 Data Complexity

Relying on claims data will give access to a limited amount of information, even though they have become very structured over time. To access a fuller picture of a health issue, clinical data in EMRs are a richer source of information as they capture more details than claims data. However, because health issues are complex and usually multifactorial, the details are so voluminous that considering them as data variables requires sophisticated analysis tools. In addition, large amounts of free-text data are included in clinical data which makes them even more complicated to analyze. Using powerful analytical tools like data mining and natural language processing (NLP) that can extract features from unstructured data and account for multiple variables during analysis helps to include the complex nature of health issues in the

measurements. This still does not resolve this challenge as the associated variables are much more numerous and from more-diverse sources than typical data mining techniques can accept while still providing high-quality models. The more variables included in analytical studies, the more each record becomes unique, and finding relationships becomes problematic.

2.1.1.6 Regulations and Requirements

Health care regulations continually evolve, and the requirements of reporting various measures keep changing. Especially when health care reforms happen and require more transparency, the burden to provide adequate reporting increases. This affects the type and amount of information needing to be captured to be able to provide the required reports. Therefore, the data collection processes and database structures require changes to accommodate the new requirements, which could cause inconsistency between, or at least challenges to deal with, different database versions.

2.1.2 Types of Health Data

Various types of health data exist, and each have their own characteristics. The data collected for research purposes – via surveys, bedside collection, clinical trials, etc. – cover a small proportion of available health data. The majority are collected routinely as part of health care delivery. The health data being collected routinely, without having any pre-defined research question, can be divided into two categories based on their purpose for collection: administrative and clinical. In operating administration systems such as insurance claims, the administrative data are generated by system users, such as professional coders and hospitals. On the other hand, clinical data are created during the conduct of medical care by clinicians or through laboratory investigations. Health claims data, primary care databases, electronic health records and disease registries are examples of routinely collected health data. Health

data may also be either routinely or non-routinely collected from other sources which introduces other types of health data, such as patient-generated data and machine-generated data [1], [21].

2.1.2.1 Administrative Data

Health administrative data are collected as a result of operating modern administration of health care. The completeness of health administrative data depends on the requirements of associated health insurance plans, operating agencies, and regulators [22]. Assessments of health outcomes and quality of care are typical secondary uses of administrative data. The sources of this type of health data include registries, life event records, insurance claims, and health service providers.

2.1.2.2 Clinically-Generated Data

This type of data is generated during clinical care, in providing diagnosis and treatment. Capture of information and input by clinicians results in deposition of data in the patient record, which can be retrieved on an ongoing basis or a per-project basis. Sources of this type of health data are electronic medical/health records (EMR/EHR) belonging to clinics, hospitals, laboratories, and pharmacies. Depending on the regulatory nature of data collection, the data may be housed in the patient record, or deposited in a data warehouse for later use in research, administration, or clinical care.

2.1.2.3 Patient-Generated Data

This type of data is directly reported by the patient to a clinician or a health care system to monitor their health status. The data collection procedure can either be directed by a health care practitioner or conducted individually by the patient. For example, the patient may decide to record this information by himself/herself for personal monitoring of symptoms, social

networking, or peer support. The sources of this type of health data include surveys, health web portals and social media.

2.1.2.4 Machine-Generated Data

A machine automatically generates this type of data, such as from computer processes and sensor signals, for passive monitoring of staff or patient behaviour and status. The sources of this type of data include remote sensors which collect or send data to a system and monitor for abnormalities. Examples include monitoring environmental exposures of a person, and remote Holter ECG monitors.

2.2 Epidemiological Studies

Epidemiology is the science of studying patterns and determinants of health-related states and events (such as diseases) in a specific population [23]. There are different types of studies that can be designed and conducted in epidemiology. They are briefly introduced as background knowledge to evaluate whether it makes sense to employ data mining techniques in all kinds of epidemiological studies.

In addition, investigating and evaluating the quality of data is one of the most important parts of any epidemiological study. Paying attention to issues like which individuals are included and missed, which variables are available, and the accuracy of the collected data plays a significant role in determining whether the results produced from the study are reliable or not. Therefore, we briefly go through the sources of bias which can invalidate the study results.

2.2.1 Study Designs

To define each of the known epidemiological study designs, first we need to understand the different characteristics that each study can have. Each study characteristic is briefly described here.

- **Type:** An epidemiological study is either "experimental" or "observational". In experimental studies (also known as interventional studies), individuals are randomly added to an exposure category (e.g., people who took the drug or did not) and are followed to monitor their outcome (e.g., cured or died), whereas in observational studies (also known as non-interventional studies), we do not have control of randomly selecting individuals in the exposure categories as they already belong to a category. Observational studies are either "descriptive" or "analytic". In descriptive observational studies, we ask the questions who, what, when, and where (e.g., who has lung cancer among people living in cold areas). On the other hand, in analytic observational studies, we ask the question why (e.g., why the risk of lung cancer increases for people living in hot areas).
- **Directionality:** The directionality of an epidemiological study is either "forward" or "backward". If we are studying people faced with the exposure (e.g., smoking) and want to see how many of them face the outcome (e.g., lung cancer) and how many did not, we are using a forward study. If we are studying people faced with the outcome (e.g., lung cancer) and want to see how many of them had the hypothesized exposure (e.g., smoked) and how many did not, we are using a backward study.
- **Timing:** Depending on when the study is done, it is either "prospective" or "retrospective". If the study is being done before the outcome occurs, it is a prospective

study. Therefore, all prospective studies have forward directionality. If the study is being done after the outcome has occurred, it is a retrospective study. Retrospective studies can have forward or backward directionality. For instance, supposed the study time has passed and we know which cases have the outcome. Either we are focusing on people with the outcome and going back to check their exposure or we are focusing on people with exposure and going forward to check their outcome.

Based on the above characteristics, there are four popular study designs in epidemiology: cohort, case-control, cross-sectional, and randomized control trial (RCT).

2.2.1.1 Cohort

Cohort studies are analytic observational studies wherein we focus on individuals having or not having an exposure and then move forward to see how many of them face a particular outcome and how many do not in each category. The study can be done either before or after the outcome is known — that is, it can be either prospective or retrospective.

The pros of this study design are that it can investigate whether the exposure is associated with the outcome, potentially the exposure's causality if Bradford Hill's criteria [24] is satisfied, and study multiple diseases at the same time. It is also low cost and quick if the timing is retrospective.

On the other hand, cohort studies are not good for very rare diseases. If the timing is prospective, the study will be costly and time-consuming (which is not good for rare diseases), and it is possible to lose track of individuals over time. If the timing is retrospective, absence of data on potential confounding factors is very likely as the data are captured in the past, and identifying an appropriate exposed cohort and an appropriate comparison group is very difficult.

2.2.1.2 Case-Control

Case-control is essentially the reverse of the cohort study design. An analytic observational study is done after the outcome has occurred — i.e., it is retrospective — by picking cases that have faced the outcome and a control group (without the outcome), and then going backward to see what each patient's exposure status was.

The pros of the case-control study design are that it can investigate multiple exposures and does not need a large sample size as the number of cases and controls are already known (i.e., it does not have the risk of proceeding with a set of individuals and none, or few, of them end up as a case).

On the other hand, the cons of this study design are that there could be mistakes in the exposure statuses if the design relies on people's memory (recall bias), it can investigate only one disease at a time, and it cannot prove the exposure causes the outcome (causality). All that can be said is that these risk factors are associated with this disease.

2.2.1.3 Cross-Sectional

A descriptive observational study on the snapshot of health experience which demonstrates the exposures and outcomes at a specific time; therefore, it is non-directional. Of course, it is done after the outcome has occurred (i.e., retrospective). For instance, we may consider the people who have ever smoked compared to people who have never smoked and check if they currently suffer from a disease or not.

The pros of this study design are that it can consider several exposures and outcomes, generate hypotheses, and measure the prevalence, and it is usually quick and cheap.

On the other hand, the cons of cross-sectional study designs are that they cannot distinguish whether the exposure preceded the outcome or the outcome influenced the exposure, they may underrepresent cases with a short duration of developing the outcome, and they only consider individuals who have survived (survivor bias).

2.2.1.4 Randomized Control Trial (RCT)

An RCT is an experimental study that randomizes people into categories of the exposure groups — without knowing who is in which category — with the intention to treat and not to harm people due to the experiment. The study goes forward to achieve the outcome (so, it is also prospective) and conclude the efficacy of an intervention by comparing the results of cases under different exposure categories.

2.2.2 Sources of Bias

Research studies in the area of health data analytics apply analytical methods, such as statistical analysis and data mining, to a data set collected in the context of health care as a matter of investigating a health issue or assisting a health care system by providing second opinions based on the past experiences. In every health data analytics study, three potential sources of error are applicable: random error, bias, and confounding. These sources of error should be addressed and verified in each study to ensure the findings' validity.

Rothman et al. [25] defines random error as that part of our experience that we cannot predict due to the existence of chance or random variations that affect the precision of measurement. On the other hand, bias is a systematic error that occurs when the investigator is designing the data analysis study and is reproducible. This error mainly happens in the data collection and information classification procedures in a particular study. Therefore, there are two major types of bias, selection bias and information bias. Confounding can be considered as another

type of bias as it also provides distorted findings systematically due to a lack of information in the study design. However, the selection and information biases cannot be eliminated from the study results if they are introduced during the study design or their possibility of happening is not reduced at the same stage, whereas results that are impacted by confounding can be adjusted or addressed. Overall, the investigator should be aware of the potential errors that are more likely to affect the validity of study findings, attempt to reduce their possibility, and report the potential impact on the experimental results.

A data analysis study can be internally valid or externally valid. If the three sources of error in a study are thoroughly investigated and addressed, then the study is internally valid. In other words, the results are accurate and reliable in the circumstances defined for that study. However, this does not guarantee the results to be true for other similar cases by just generalizing the conditions. For example, after handling the sources of error, the results of analyzing data collected from a population with asthma located in a specific region are only valid for the people living in that region. The findings would not necessarily be applicable to the population residing in another region. The external validity of these results can only be claimed when they are properly discussed and generalized based on rational facts and similarities.

It is possible that the selected analytical method opens the door to different set of potential errors. For instance, if an analytical method requires sampling in various stages to conduct the analysis, the chances of introducing selection bias to the study due to using a non-random sampling method increases.

When reporting the results of a study, the risk of potential errors (such as random error, bias, or confounding) should be included so that the reader can conclude whether the results and findings of the study are valid and acceptable or not.

There are many types of bias included under selection bias and information bias, and not all of them are relevant to our discussion in this research study. In the following paragraphs, we will briefly introduce the different sources of bias and confounding, and then discuss the specific types that are expressed during this research study.

2.2.2.1 Selection Bias

Selection bias refers to bias in the analysis results that comes from the procedures that select the individuals into the study (i.e., that form the study population) [26]. It can be introduced during both the design and implementation of the study. Limitations in collecting or accessing the appropriate data contribute to selection bias. As an example, if we only have access to the data of individuals covered by public health plans, then any case from the source population that does not have a public plan is simply missed, and the research conducted using the selected sample is at risk of bias because people covered by the public health plan may be systematically different from people using only private insurance (in socioeconomic status, health services utilization patterns, etc.). In addition, we often select a sample from the larger population in statistical analysis. The criteria for including and excluding sample cases determine whether a selection bias is systematically injected into the study design. Selection bias also occurs if the distribution of variables individually and in each group are not equivalent between the study population and the source population. For instance, if 30% of the source population suffer from the concerned disease, the same proportion should exist in

the study population, and if 20% of the diseased population are exposed to a factor, then the same proportion should, again, exist in the diseased sample for the same factor.

Selection bias can be discussed further and elaborated on how it is more likely to be introduced to the study in different study designs. For example, if it is a case-control study, then the distribution of exposed cases, exposed controls, non-exposed cases, and non-exposed controls in the study population should be the same as (or at least close to) their distribution in the source population. However, we have covered the main aspects of selection bias. Below, we briefly present some specific types of selection bias that will be relevant to this research study and addressed in the Contextual Preprocessing Guideline in section 5.2.

2.2.2.1.1 Ascertainment Bias

With ascertainment bias, the selected individuals in the study population are not equally representative of the source population [26].

2.2.2.1.2 Inclusion Bias

Inclusion bias occurs in case-control studies where the number of exposed controls in the study population is higher compared to the source population [25], leading researchers to underestimate the association between the exposure and outcome.

2.2.2.1.3 Exclusion Bias

Exclusion bias happens in case-control studies where the number of exposed controls in the study population is lower compared to the source population, while the exposed cases remain the same [27], leading researchers to overestimate the association between the exposure and outcome.

2.2.2.1.4 Matching Bias

Matching bias occurs in case-control studies where the controls are selected by matching with the cases based on a set of definite variables. This can lead to overmatching when the matching is done by a non-confounding variable — associated with the exposure, but not with the outcome — and the association is underestimated [25].

2.2.2.1.5 Non-Random Sampling Bias

Non-random sampling bias refers to cases where the individuals in the study population are not selected randomly and produce a non-representative sample [28]. Therefore, particular segments of the source population could be missing, causing distortion in the distribution compared to the source population.

2.2.2.1.6 Survivor Treatment Bias

Survivor treatment bias appears in retrospective observational studies where the cases who live longer — due to other factors — get more chance of receiving the treatment, leading to overestimating the association between the treatment and survival [29].

2.2.2.2 Information Bias

As discussed earlier, selection bias could exist in data if the correct patients or individuals are not selected and included in the data set. In addition, the accuracy of values in the collected data are also important. If the values are not accurately captured and presented, information bias is introduced to the study. Many different reasons could cause these errors to happen, such as incomplete medical records, recall mistakes in reporting a value in a questionnaire, and data entry errors.

Below, we briefly present some specific types of information bias that will be relevant to this research study and addressed in the Contextual Preprocessing Guideline in section 5.2.

2.2.2.2.1 Misclassification Bias

Most issues yielding information bias are related to misclassification of values. Therefore, the terms “information bias” and “misclassification bias” are sometimes used interchangeably. If there are errors in the status of exposure or other input variables, the association between those variables and the outcome or target variable is over- or under-estimated. Random errors usually lead to misclassification bias, such as random errors in capturing the data or having missing data. Furthermore, different clinical approaches in diagnosis could results in different diagnosis codes that might not correctly represent the health issue in the patient.

Misclassification bias is a serious source of bias in studies using routinely collected data, which includes a remarkable portion of health administrative data, since they heavily rely on coding which may not be accurate or complete. The information can come from different sources or groups with different attention to completeness and accuracy. Misclassification in the exposure or input variables could happen due to many different reasons; therefore, there are various types of misclassification bias. If the misclassification occurs only in one of the comparison groups, this is differential misclassification; whereas, if the misclassification exists in all groups, then it is called non-differential misclassification. For instance, in collecting data to study a disease, the values of an exposure are captured from the tests of patients suffering from the disease who visited a hospital as cases, while the same information for members of the control group (the ones who are not suffering and therefore did not visit the hospital) are gathered from another source. This could potentially lead to differential misclassification bias.

2.2.2.2.2 Immortal Time Bias

Immortal time bias refers to cases where the exposure did not have enough time to possibly occur and the patient (case or control) receives credit for being at risk while the required length of follow-up was not met [25]. Therefore, the exposure status might be incorrect and lead to invalid — risk or protective — association with the outcome.

2.2.2.2.3 Publication Bias

Publication bias occurs when only the studies or results that match the positive findings or preconceived opinions of authors, editors, reviewers, and organizations are published, instead of all related studies or all valid results found in the study being published, including the negative or perceived less important ones [26]. While this bias is more important for RCT studies, it is applicable to all research studies.

2.2.2.3 Selection and Information Biases

There are some types of bias that, depending on how the study is designed and conducted, can occur as a source of selection bias (in selecting cases) or information bias (in constructing features). If the cases are being selected based on a set of criteria to be included or excluded from the dataset, then these types of bias are acting as a kind of selection bias. On the other hand, if a variable representing an exposure or outcome is being constructed to demonstrate whether the individuals included in the dataset have any history of a certain disease based on a set of criteria, then these biases will happen as information bias.

Below, we briefly present some of these sources of bias that will be relevant to this research study and addressed in the Contextual Preprocessing Guideline in section 5.2.

2.2.2.3.1 Prevalence-Incidence (Neyman) Bias

Prevalence-incidence bias, also known as Neyman bias, occurs in both cross-sectional and case-control studies where only a series of survivors are selected [30] and the exposure influences mortality due to the outcome, which leads to underestimating the association [31]. In other words, this bias will likely occur when prevalent cases (active cases in the population) are selected instead of incident cases (all cases in the population during the time window of study).

2.2.2.3.2 Diagnostic Bias

Diagnostic bias is where the patient is or is not diagnosed with a disease by mistake due to many different reasons such as having a long latency period. The diagnostic bias can be expressed more specifically as well. For instance, "purity diagnostic bias" refers to the situation where cases are excluded due to having other comorbidities [32] and "diagnostic suspicion bias" occurs when the knowledge of other general factors (e.g., ethnicity) of the patients affect the intensity or outcome of diagnosis [33].

2.2.2.4 Confounding

Confounding happens when the exposure or an input variable is correlated with the outcome or target variable, but another variable acts to modify both the likelihood of having the exposure and the outcome. That variable is a confounder. For instance, alcohol is associated with lung cancer, but alcohol consumption is not causally related to developing lung cancer. Alcohol consumption is an intermediary variable because smoking increases the chance of consuming alcohol and is the main reason for developing lung cancer at the same time. In studies using administrative data, this error can be expected a lot as the clinical factors do not exist in the data (a situation referred to as unmeasured confounding).

2.3 Data Mining

In the middle of 1990s, data mining came into existence as a strong tool to extract useful information from large datasets and find the relationship between the attributes of the data [34]. Data mining originally came from statistics and machine learning as an interdisciplinary field [35]. It then grew in applicability to the point that in 2001 it was considered as one of the top 10 leading technologies which would change the world [11].

2.3.1 Main Techniques

Data mining techniques are divided into two main categories: descriptive (or unsupervised learning) and predictive (or supervised learning) [12][13][36][37]. Descriptive data mining is an exploratory analysis that attempts to measure the similarity of records and discover the patterns and relationships. The most important techniques in descriptive data mining are clustering and association. On the other hand, predictive data mining tries to generate predictive rules as a model to classify the records based on a specific target (or label). Classification is the most widely used technique in predictive data mining. As there is no standard framework for a data mining process [38], it is important that the analyst holds strong skills in this area and understands the techniques very well to design an appropriate framework and achieve high quality and reliable outcomes. The main techniques have been introduced briefly in continue.

2.3.1.1 Classification

This technique is used when the data are required to be classified into different groups based on a target attribute – e.g., patient cost. Therefore, the classifiers predict the target label for each record using the input attributes. Some of the most widely used classification techniques are decision trees, neural networks, K-nearest neighbors, support vector machines, and

Bayesian methods. According to a survey [35], decision tree algorithms are the most popular ones among all other classification techniques in the applications of data mining to health care.

Classification techniques are widely used in health data analyses, including: analyzing microarray data [39], diagnosing skin diseases [40], performance of different classifiers on cancer datasets [41], predicting cost of health care services [42][43], identifying significant factors in health care coverage and predicting the status [44].

2.3.1.2 Clustering

This technique is used when we do not have much information about the different types of data objects involved in a population. As it is an unsupervised learning, it tries to find the cluster of data objects that have similarities to each other without considering any specific target label. Therefore, there are no predefined classes in contrast to classification. Different clustering techniques are partitioned clustering, hierarchical clustering, and density-based clustering.

2.3.1.3 Association

This technique is used when the relationship of attributes in a dataset needs to be identified – e.g., the association among the purchased items of a customer's basket. This technique is mainly used in health care to detect the relationship between diseases. In addition, association techniques can also be combined with classification techniques to increase the capability of analysis. For instance, the rules in a database or relationship of attributes in a dataset are detected, and then an efficient classifier is built by just considering the identified rules and including just the main attributes.

2.3.2 Relationship with Epidemiological Study Designs

The main goal of this research study is to integrate data mining with epidemiological studies. Now that we have reviewed the main techniques in data mining in addition to epidemiological study designs, we digest the relationship between these two topics. Which epidemiological study designs are relevant when using data mining techniques? We briefly discuss this below under each type of data mining method.

2.3.2.1 Predictive Modelling

In this method of analysis, we have a target variable and need to train a classifier to build a predictive model. Therefore, the data from past events and their outcome should exist. This means that the type of study is "observational" and the timing is "retrospective". Typically, our focus is on the target, and we consider the distribution of classes, then we look back to investigate which exposures have higher association with the outcome. So, the direction is "backward". Putting all this together, a data mining study that attempts to build predictive models most likely requires a "case-control" study design from an epidemiological perspective.

These studies discover rules and relationships and normally generate new hypotheses. To test these hypotheses, "cohort" studies can be conducted using the traditional epidemiological approach to verify them.

2.3.2.2 Descriptive Modelling

Even though we attempt to describe the characteristics of data, descriptive modelling is more than just descriptive epidemiology — i.e., extracting distributions in variables, certain regions, or a specific period. Techniques included in descriptive modelling can find association rules

between the input variables and discover clusters of people without any assumptions or predefined segmentations like regions.

The data are already collected, but no target variable exists or is included in modelling. This means that the type of study is “observational” and the timing is “non-directional”. As the main attempt is to describe and analyze the current status, a descriptive modelling study most likely requires a “cross-sectional” study design from an epidemiological perspective.

Like predictive modelling studies, these studies also normally generate new hypotheses. And to test these hypotheses, “cohort” studies can be conducted using the traditional epidemiological approach to verify them.

2.3.3 Advantages over the Traditional Statistics

In the recent history, traditional statistics has been considered as the main data analysis method and still actively contributes to most studies. Although statistics are viewed as the *primary* data analysis in the current era, data mining is considered as the *secondary* data analysis due to its strengths and rapid developments [45]. While the fundamentals of both analysis methods are mathematics, data mining does include statistical analysis as part of its process. However, as an interdisciplinary field which benefits from the advantages of other fields, such as machine learning, artificial intelligence, and visualization, it has some important gains over the traditional statistics.

First, statistics prefers to use more conservative strategies in the first phases of analysis, and in general, employ concrete mathematical methods to run analysis, and provide crisp and exact results. On the other hand, data mining is open to consider various approaches in regards to

mine the data in different orders [46]. Due to this flexibility, data mining uses heuristics as well when facing with real-world issues and offer fuzziness in the outcome results.

Second, although data mining has an inherent relationship with statistics, its techniques are designed to perform faster on large data sets – over millions of data points. Scaling up statistical methods is a challenge as a lot of them have high complexity in computation and running them on distributed systems requires new algorithm designs and tuning to reduce the computational cost. This challenge becomes even more serious in real-time analyses where fast responses are essential [10].

Finally, statistics conducts confirmative analyses where a hypothesis is tested – proved or rejected (hypothetico-deductive analysis). On the other hand, data mining studies are explorative and do not consider any clear hypotheses, i.e. the data are explored and new information is retrieved (inductive analysis) [46]. In other words, statistical analysis methods are hypothesis-driven, whereas data mining methods are discovery- or data-driven. This advantage can be particularly useful when studying the prevalence of new diseases with unknown root factors.

2.4 Data Mining in Health Care

After reviewing the basic information, we discuss some advanced topics on the relationship between data mining and health data, including applications and challenges. Finally, we further investigate which applications benefit more from boosting the analytical platform.

2.4.1 Applications of Data Mining to Health Care

The application of data mining to health care can be categorized into four groups: clinical decision making, population and public health, genetics and biomedicine, and health

administration policies [19]. Although “text mining” has important health care applications, it is excluded from the mentioned categorization as it is considered to be a separate and broader field of research with its own set of applications to health care. It focuses on natural-language processing techniques and uses data mining as one of many components.

The categories and their descriptions are retrieved from past systematic reviews. Each is briefly introduced, along with some examples.

2.4.1.1 Clinical Decision Making

Applying data mining techniques at a clinical level can help to predict high-risk patients, and provide additional information to clinicians. As an example, Healthways’ data mining software discovers relationships to predict patient risk and suggests appropriate intervention and prevention plans based on the available evidence [47]. Some new studies are also focusing on data stream mining to support real-time clinical decision making which requires very high performance data mining algorithms [48].

2.4.1.2 Population and Public Health

Patterns, trends, and causes of disease can be investigated using data mining, toward enhancing public health awareness and improving health status of the population. For instance, breast cancer patients have been classified and investigated epidemiologically using data mining techniques [41]. In addition, data mining techniques can help to analyze the survivability and identify the significant factors affecting survival from specific diseases, including breast cancer [49] and kidney dialysis patients [50].

2.4.1.3 Genetics and Biomedicine

In contrast to studies done in the health of populations, biomedical data analysis research investigates diseases at the molecular level. Supporting data are retrieved from genomics, transcriptomics, and proteomics of humans and microbiota. Profiling gene expression in the diagnosis of leukemia is an example of applying data mining on human DNA microarray data [51]. Also, researchers can classify the microarray data of diseases [39][51] for different purposes such as predicting whether a patient will have a health condition [52], and differentiating similar diseases based on their DNA expression microarray results extracted from the infected sample cells or tissues [53].

2.4.1.4 Health Administration Policies

Health administration systems are operated by health care providers, insurance companies and public-sector institutions. Analyses are used to assess the accuracy and efficiency of their procedures and policies. A well-known study type in this category is insurance fraud detection using data mining [54]. In addition, as the health care cost is increasing globally, some health administrative data analyses are mainly focused on identifying high-cost patients, and solutions for preventing significant costs by providing early stage care if possible [43].

2.4.2 Necessity of Data Mining in Health Care

In addition to the items mentioned in the "Applications of Data Mining to Health Care" section, such as fraud detection, causing factors discovery in unknown diseases, microarray data classification, and high-cost or high-risk patient prediction, that makes data mining techniques necessary in healthcare, health data analysis is going toward a direction that a tool like data mining is becoming more important and essential to provide successful analyses. The data being collected in the health sector have their own specific characteristics that makes any

analysis very challenging. Some of the main concerns in this regard has been described in the next section. However, there are other specifications tied to health data that prevent traditional methods to evolve and provide new information.

2.4.2.1 Health Big Data

"Big Data" has become a known term referring to a massive collection of data to the point that the traditional data analysis techniques do not function efficiently. In addition to the volume of data, velocity (streaming data) and variety (semi-structured and un-structured) are major concerns in "big data" too. As a result of implementing EHR and similar systems to collect electronic health and medical data, in addition to the data being collected via online insurance claims and health surveys, a large amount of data has been collected by health organizations which is increasing on a daily basis. Data mining techniques have the ability to consider the whole population into the analysis which can provide information about specifics, details, and also minorities.

Although data mining algorithms also need to be improved in terms of performance and scalability, at least current data mining algorithms have the capability of working with large datasets. Moreover, new frameworks are being studied and developed to handle health big data analytics, in terms of better understanding the data, overcoming with the challenges, and providing higher quality results. Normally, the steps that need to be taken to support health big data analysis are: data aggregation, data maintenance, data integration, data analysis, and pattern interpretation [55].

2.4.2.2 Health Condition Mysteries

As discussed earlier, not a lot of information is available about genes' impacts on health disorders. Therefore, exploratory methods can significantly help the researchers to identify the

relationships between the attributes and also cluster the groups of contributing genes. Descriptive data mining has started evolving in this area with the attempt of going from specific (data) to general (knowledge) about the effects of genes on health conditions. In the next steps, classification and prediction models can also be built for the discovered groups to be able to predict the health disorders using gene expression patterns observed in patient's related organs.

In addition, by the changes introduced in human's lifestyle and living conditions in the modern era, some known rare diseases start to become more prevalent such as Inflammatory Bowel Disease (IBD). On the other hand, new viral infections occur and spread globally, such as Ebola and COVID-19. The main causes of these diseases and dysfunctions may not clearly be identified at the time of outbreaks and even until now, and based on the previous observations, more of these new diseases will be discovered in future as well. In this case, data mining techniques can also be used to discover the trends related to these diseases considering clinical, biomedical, and even environmental factors.

2.4.3 Challenges of using Data Mining in Health Care

There have been some studies conducted to examine the issues and challenges of applying data mining techniques in healthcare [56][57]. In overall, the top important challenges that make data mining in healthcare (and biomedicine) applications differ from other application are as following:

2.4.3.1 Data Quality

To achieve useful and reliable information from a data mining process, it requires data with good quality. Health data are usually collected from different sources with totally different set-ups and database designs which makes the data complex, dirty, with a lot of missing data,

and different coding standards for the same fields. For instance, although problematic handwritings are no more applicable in EHR systems, the data collected via these systems are not mainly gathered for analytical purposes and contain many issues – missing data, incorrectness, miscoding – due to clinicians' workloads, not user-friendly user interfaces, and no validity checks by humans. In addition, as EHR data are observational and not experimental, they might not represent all cases involved in an issue and face with different types of biases (selection, confounding, measurement) which makes it challenging to generalize the outcomes to the whole population [58]. Dealing with this type of data is not easy and requires significant efforts in cleaning and pre-processing before building the actual data mining models. Moreover, there have been some studies to compare and examine the difference of EHR data with other types of data collection methods, such as health surveys, in analyzing specific health conditions. For instance, on study [59] identified that the prevalence of self-reported multimorbidity was significantly higher in health survey data – compared to EHR data – among younger patients, while the prevalence was similar in both data sources for elderly patients. Also, self-reports appeared to be more sensitive to identifying symptoms-based conditions.

If a health service provider – e.g., a hospital – wants to gather data so that they can be used for secondary purposes, especially analytics, to obtain useful information for improving their service quality and making their costs efficient, it is essential to define frameworks and methods that data with acceptable quality be collected from the first place. This requires very flexible and scalable database designs for every unit that can easily be linked – whenever required – while using the same standards in entering the data. For instance, when the patient has to fill out a lot of questions at the check-in time when visiting a hospital, he/she is sick

and might not even have enough concentration in answering the questions; therefore, this will lead to collecting data with lots of missing and even incorrect entries. Other considerations which can make a gathered dataset – e.g., via EHR systems – appropriate for secondary uses is to have interoperable data collecting systems with improved infrastructure, and also using proper data analysis techniques [58].

2.4.3.2 Data Sharing and Privacy

Since the health data contains personal health information (PHI), there will be legal difficulties in accessing the data due to the risk of invading the privacy. Therefore, data analysts can not easily access the collected health data. Health providers are not legally and ethically permitted to share their data with the analysts to avoid any risk that threats the privacy of patients. On the other hand, preparing a secure infrastructure to gather data from different sources is very time consuming and expensive [38]. No access to data essentially means no input into the data mining techniques, and subsequently, no analysis and outcome information.

To be able to access the required health data for analysis, there should be sufficient security protocols implemented in the data warehouse, so that analysts can reach the data to run the required analysis. Although implementing these security protocols impede the speed and increase the cost of accessing these data, they are important to have them in place to both protect the privacy of patients as much as possible, and allow the analysts to access data to be able to conduct analysis. In addition, health data can be anonymized using masking and de-identification techniques, and be disclosed to the researchers based on a legal data sharing framework [60]. However, the amount of anonymization is very crucial. If the data get anonymized so much with the aim of protecting the privacy, on the other hand they will lose their quality and would not be useful for analysis anymore. Therefore, coming up with a

balance between the privacy-protection elements (anonymization, sharing agreement, and security controls) is essential to be able to access data that are usable for analytics.

2.4.3.3 Relying on Predictive Models

There should not be unrealistic expectations from the constructed data mining models. Every model has an accuracy. When a predictive model for diagnosing a health issue – e.g., diagnosing the type of thyroid cancer – is built, it usually does not have an accuracy of 100%. Depending on the amount of model's accuracy and the importance of the decision being made using that model, we should decide how much we can rely on the outcomes of this data mining model. Especially when we are dealing with clinical medicine decisions based on the outcomes of one or multi data mining studies, it is important to consider that it would be dangerous to only rely on the predictive models when making critical decisions that directly affects the patient's life, and this should not even be expected from the predictive model.

However, data mining models and specifically predictive models, could be very helpful in flagging potentially at-risk patients or providing a second opinion for the physician's decision on a treatment based on previous cases and decisions. In some cases, the doctor might have under-estimated the severity of a patient's disease, and the constructed model will alert the doctor to consider some aspects that might have been missed and makes the patient to be at risk. On the other hand, if the doctor is not too much sure if for instance a diagnostic surgery is required for this patient, then predictive model will help the doctor to decide – either with going for the diagnostic surgery or not – with more confidence by considering the previous cases via the constructed model.

2.4.3.4 Variety of Methods and Complex Maths

As the underlying math of almost all data mining techniques is complex and not very easily understandable for non-technical fellows, thus, clinicians and epidemiologists have usually preferred to continue working with traditional statistics methods as it is easier to communicate its technicality. They generally understand the *p* value better comparing to data mining's measurement methods such as correctness and ROC curve. In addition, as there are many different data mining techniques and methods, it is difficult for clinicians to get familiar with all the different methods and easily select the proper one.

It needs to be considered that usually there is no one best technique that works for different datasets. It is essential for the data analyst to be familiar with the different techniques, and also the different accuracy measurements to apply multiple techniques when analyzing a specific dataset. This way, the chance of getting better results and outcomes will be higher. In this case, it would be necessary to get help from analysts that are more familiar with the different data mining techniques to be able to apply different methods and take advantage of data mining's strength in dealing with large datasets and accepting more input attributes comparing to traditional statistics.

2.4.4 Big Data Analytics in Health Data Mining

The approach chosen for this section is to categorize and investigate the applications of data mining to health care using past systematic reviews. Therefore, each category represents an application of data mining to health care. One or more sub-categories are defined under each category. Each sub-category represents a set of studies with common characteristics from a data analysis perspective in that category. In other words, studies in each sub-category share similar technical concerns, limitations, and possible solutions in analyzing their data. For

instance, two studies might not have much medical relationships with each other and still both fall into the same sub-category, while two other studies focusing on the same health condition have different data analysis characteristics. The available types of health data for particular applications are determined, as well as which dimensions of big data analytics (the four V's) are applicable to leverage analyses for each sub-category. Eventually, the possibility of enhancing the data analysis quality using big data analytics for each group of studies is assessed.

The collected information is presented and summarized in a hierarchical, tree-style format in Figure 2.2. The horizontal levels of the tree represent (1) the main categories of applications; (2) the sub-categories; (3) the associated health data types for the sub-category; and (4) the applicable dimensions of big data analytics for the same sub-category.

As an example, "Population and Public Health" is a category of data mining applications to health care, and studies with "Responsive Conditions" are grouped as a sub-category. For these particular studies, patient-generated data could be a source of information. Since the concerned conditions in this group are usually spread quickly among a population and/or require real-time analysis, the velocity of fetching data could be a limitation in running high quality analyses. This limitation can be mitigated using a big data analytics environment.

In continue, first we introduce the dimensions of big data, and then discuss our investigations on each application of data mining to health care to introduce the known group of studies in each application, present the available health data and argue which big data dimensions are applicable.

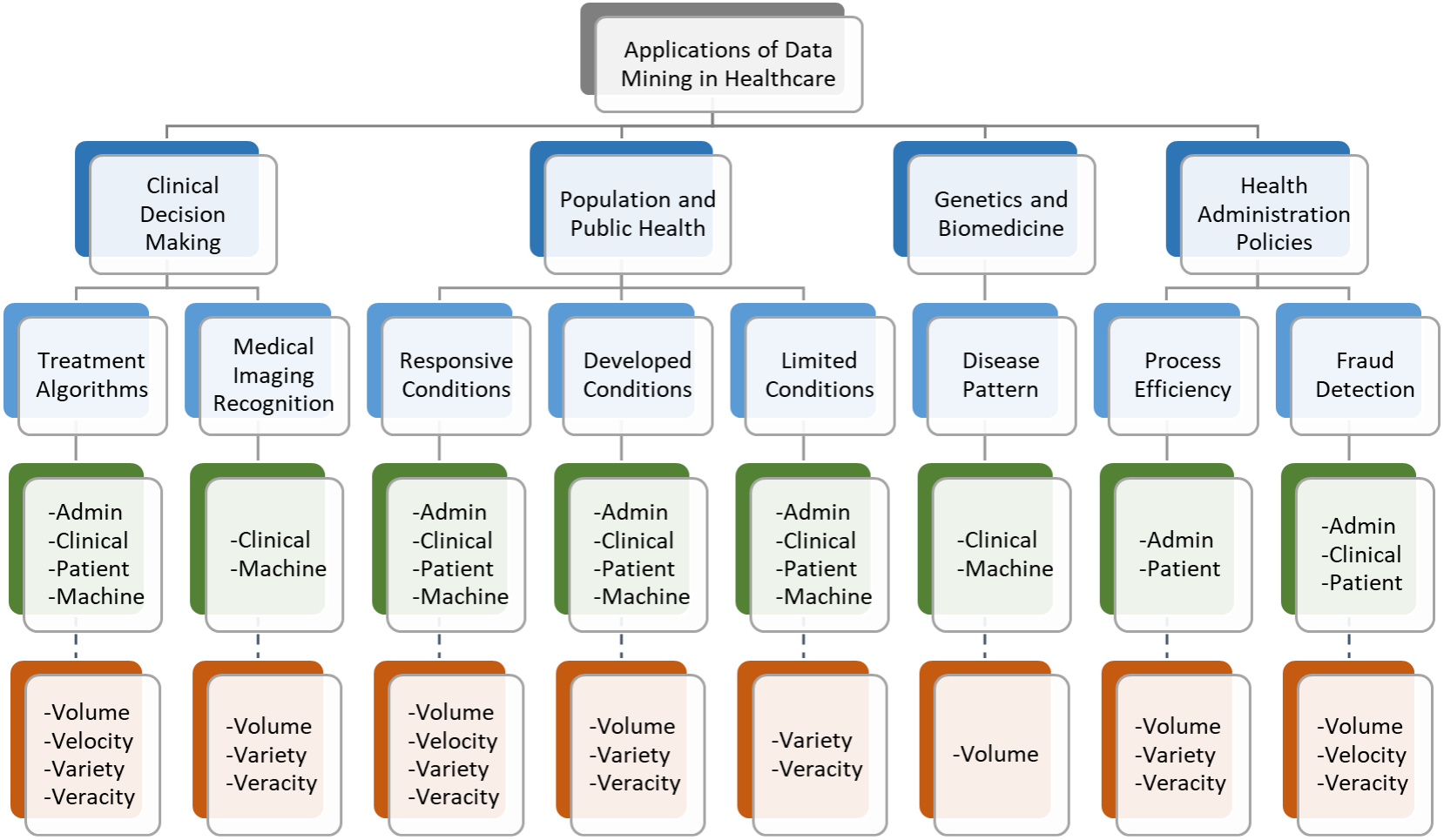

Figure 2.2: Big data dimensions and accessible health data for applications of data mining in health care [18]

2.4.4.1 Dimensions of Big Data

Enormous amount of data of many different types and conditions are being collected on a regular basis. This includes streaming modes in various industries and sectors. In modern health care, due to the transition from paper-based systems to digitized systems using electronic means, large amounts of data are generated, often continuously. These recorded data are of different quantities, formats, and conditions. Therefore, health data may qualify as being big data.

To evaluate this, the characteristics of a big data need to be understood. Although an evolving concept, "big data" refers to data that are large in volume, varied in type, fast in generation velocity, and of uncertain veracity (accuracy) [61]. Each of these dimensions can be defined briefly as following:

- **Volume:** Big data are large in scale (e.g., petabytes)
- **Velocity:** Big data are generated at a fast pace and often in real-time (e.g., streaming)
- **Variety:** Big data contain different forms of data – structured, semi-structured and unstructured
- **Veracity:** Big data can lack high quality, and have uncertain accuracy

Characterized by these four V's (and sometimes more), big data can be defined as a collection of data so large and complex that it becomes difficult to process using existing data analytics techniques. Analysis of big data relies on scalable distributed platforms, such as Hadoop and Spark, which support MapReduce structure optimized for analysis of massive data in parallel.

2.4.4.2 Clinical Decision Making

Studies in this category are mainly at the clinical level. In practice, most of these studies are done either to develop treatment algorithms and clinical pathways, or to help recognize specific issues and patterns in medical imaging artifacts [62]. Suitable types of health data for these studies and applicable dimensions of big data are discussed for the two sub-categories.

2.4.4.2.1 Treatment Algorithms

In clinical care, developing a clinical pathway is typically a sensitive procedure, requiring attention to ethical considerations, as it deals with human lives. A clinical pathway or algorithm could be a stream of investigation, diagnosis, treatment, or some combination. Data mining techniques are employed to provide additional information in designing these algorithms, based on experiences. In addition, clinicians would avoid under-estimating the effects of a symptom, and make treatment decisions with higher confidence by receiving supportive second opinions using analytical model. This could eventually reduce unnecessary costs for both care providers and insurance companies. Studies that attempt to predict high-risk outcomes for patients based on their symptoms and individual determinants fall in this category [63]–[66]. Data gathered by practitioners and stored in EMRs from examinations and lab tests, data reported by the patients themselves (either with directions from a clinician or not), and even data generated by health monitoring devices are potential sources of information to conduct these studies.

The number of EMR vendors is rapidly increasing, as is the choice of modules available within each system. Developing treatment algorithms to provide higher quality clinical service could be acute and time-sensitive, which would require analyzing streaming input data. Sources of information can involve both structured and unstructured text materials, including physician's

notes and prescriptions. The collected data typically come from different sources and individuals, which could introduce conflicts and inaccuracies within the data.

Since the available data for this set of studies contain all dimensions of big data – volume, velocity, variety, and veracity – the possibility of enhancing the quality of analysis using big data analytics is "very high".

2.4.4.2.2 Medical Imaging Recognition

Examples of this set of studies include diagnosing melanoma by analyzing digitized images of skin lesions [67], identifying minimal hepatic encephalopathy in patients with cirrhosis based on imaging of brain white matter [68], discrimination of benign and malignant breast nodules using multiple ultrasonographic features [69], and differentiating neoplastic and non-neoplastic brain lesions based on magnetic resonance spectroscopy [70]. Data mining studies on anomaly detection in high-resolution high-dimensional medical imaging data which require extensive processing capacities also fall in this category.

The two main sources of information here are (1) the data gathered by clinicians as a result of examining patients prior to sending them for lab work (clinically-generated health data); and (2) the data generated by medical imaging machines in screening organs under study (machine-generated health data).

The quantity of screenings done using medical devices in health centers is usually high, supplementing clinical information gathered by clinicians. In addition, medical images are typically translated into high-dimensional data. The format of collected data includes images, videos and unstructured text, as well as other structured data. The collected data could include screening errors and vague images, increasing uncertainty of the data quality.

The possibility of enhancing the quality of analyses using big data analytics for this set of studies is "high" as the available data contain at least three dimensions of big data – volume, variety, and veracity.

2.4.4.3 Population and Public Health

This type of research aims to determine the causes, trends, and patterns of diseases (population health), and to analyze the practice of medicine associated with diseases (public health) at the population-level. The goal is to improve the overall health of the public. Depending on what type of disease and under which condition is being analyzed, the requirements of a study might change. Therefore, they have been divided into three main groups, based on the commonality of their characteristics from a data analysis perspective: responsive conditions, developed conditions, and limited conditions. Note that this categorization is not classifying the diseases from medical point of view. The group naming is done based on sharing similar conditions and requirements to analyze data. Suitable types of health data for each of these studies and the applicable dimensions of big data are discussed following.

2.4.4.3.1 Responsive Conditions

Fast-spreading diseases usually infect a large proportion of the population in a short time, typically caused by viruses (e.g., H1N1, Ebola and influenza) or bacteria (e.g., *C. difficile* and *E. coli*). To control the spread, real-time analyses are conducted to anticipate the next at-risk regions. In overall, studies analyzing conditions that spread across or affect a proportion of population which require quick awareness and analysis fall in this group. Therefore, these studies are not just limited to controlling and monitoring fast-spreading diseases, and includes any high incidence conditions that need rapid response. Another example is predicting the number of asthma-related emergency department (ED) visits in a specific area in near-real-

time using environmental and social media data to help public surveillance, ED preparedness, and targeted patient interventions [71].

Data sources of this set of studies accumulate information from patient visits to healthcare centers (producing administrative data), examinations by clinicians and undergoing lab tests (creating clinically-generated data), sharing of their experiences and symptoms on social media to communicate with experts or other affected people (developing patient-generated data), and also environmental data collected by sensors as seen in the latest example (gathering machine-generated data).

When an outbreak occurs, such as the recent Zika virus in Brazil [72], large amounts of information and records are typically gathered. It is important for public health practitioners and the health system to react quickly and analyze data immediately, and as they change in real-time. The collected information could include both structured and unstructured materials such as experts' notes, affected people's video messages, and formal record capture. Usually the cause is unknown initially, and patient symptoms could overlap with those of other health issues; this can lead to capturing uncertain, irrelevant, and even incorrect information.

In this sub-category, the available data contains all dimensions of big data – volume, velocity, variety, and veracity; hence, the possibility of enhancing the quality of analysis using big data analytics is "very high".

2.4.4.3.2 Developed Conditions

This group includes studies analyzing conditions that do not require real-time processing – as the critical stage of disease development has passed – or even cannot be processed instantly – due to the nature of disease which needs to develop to a certain point to provide further details.

In addition, these conditions usually hit a large proportion of the population and typically have high prevalence. For instance, chronic diseases are not passed to other people and might not have the same incidence rate as the fast-spreading diseases, but they often cause severe conditions and last longer in a population. The incidence of breast cancer, as an example, started to grow in most societies and now is still considered as a prevalent issue among a significant portion of global population which there were attempts in classifying and predicting the suffering patients [41]. Another example is identifying the contributing factors of developing asthma in a population, from which more than 230 million people suffer worldwide, and is considered one of the most common chronic diseases among children [73].

Note that not all studies that focus on chronic diseases fall in this sub-category. As mentioned in the previous section, the study on anticipating the number of asthma-related emergency department visits [71] is considered a "responsive condition" as requires real-time analysis at population level. On the other hand, the study that predicts asthma attacks of the patient by analyzing sensor data [74] is similar to Treatment Algorithms of Clinical Decision Making studies. Therefore, the category of every health data analysis study is determined based on its characteristics.

For analyzing this type of conditions, administrative health data can be used to track the development of the disease, in addition to clinical data provided by practitioners or patients themselves. Data generated by medical imaging devices for screening the involved organs can be considered as well.

Due to the high prevalence of the concerned condition in the focused population, the volume of data collected from both administrative and clinical perspective is large. However, in these

cases the speed of generating and collecting data might not be high, comparing to the responsive conditions. The format of the collected data can be diverse, as well as the quality.

The available data for this set of studies contain at least three dimensions of big data – volume, variety, and veracity –, therefore, the possibility of enhancing the quality of analysis using big data analytics is "high".

2.4.4.3.3 Limited Conditions

In this group, diseases have long duration and often unknown causes, hitting a smaller proportion of a population in contrast to the "developed conditions". Most immune-mediated diseases comply with this condition (e.g., inflammatory bowel disease and multiple sclerosis) in many societies; however, not all of them usually fall in this group (e.g., asthma). Note that a disease can be rare in one population, while quite common in another. No matter if the disease is considered rare in the focused population or not, as long as the amounts of collected data to analyze the disease at a population level are no more than few millions of data cells of an ordinary structured data (e.g., hundreds of thousands of rows with tens of columns), the study is considered as "limited condition" in this categorization. There are generally attempts made in analyzing and investigating these diseases to narrow down the contributing risk factors. Although some genetic diseases may be rare, their root causes have usually been identified; therefore, the related analyses can focus on genetics data, which are introduced in the next category of applications. The sources of information for this group of studies can be the same as the previous group, as patients usually go through the same procedures: administrative data, clinically generated data, patient-generated data and machine-generated data.

In contrast to responsive conditions and developed conditions, studies dealing with limited conditions do not have the luxury of access to large amounts of data. Similar to the developed conditions, the input data are not collected at high rate. However, the format of the collected data can still be diverse, and the quality.

The possibility of enhancing the quality of analysis using big data analytics is “medium” for this sub-category as the available data contain at least two dimensions of big data – variety and veracity.

2.4.4.4 Genetics and Biomedicine

Research in this category focuses on disease caused at the molecular level, using data retrieved from genomics (DNA molecules), transcriptomics (RNA molecules), proteomics (proteins), and metabolomics (metabolites) of human hosts and their resident microbiota. These aspects can be used in determining an individual patient’s unique biological profile toward applying the most effective therapy, also known as personalized medicine. The knowledge of gene and protein expression procedures has improved rapidly over recent years. Therefore, the amounts of prepared microarray data have increased also. These developments allowed researchers to analyze health issues at the microscopic level, and to determine relationships among different aspects of the biology of humans and disease. Data mining is already being widely applied to microarray data, mainly for detecting patterns of diseases. Suitable types of health data for each of these studies and the applicable dimensions of big data are discussed below.

2.4.4.4.1 Disease Pattern

Examples of this form of research include profiling breast cancer subtypes based on pervasive differences in their gene expression patterns [75], and improving the prediction of colorectal cancer using gene expression signatures [52]. Initial data gathering is done in laboratories, by

extracting genomic DNA samples (clinically-generated data). DNA segments are then sequenced by computer processes (machine-generated data), to provide microarray data after further procedures that demonstrate the expression level of genes.

Although the process of sequencing the human genome can now be done faster with much lower cost, the whole genome sequence consumes hundreds of gigabytes of storage space, and can require significant processing time and cost [76]. When a disease with a known cause is being studied, only a specific part of the genome may need to be sequenced for analysis. However, for unknown diseases, the whole genome (or even the whole exome, including non-coding parts of the genome) may be required, and the amount of data produced becomes massive. On the bright side, typically whole genome sequencing is now done with high accuracy.

As the available data for this set of studies contain at least one dimension of big data – volume –, the possibility of enhancing the quality of analysis using big data analytics is "low".

2.4.4.5 Health Administration Policies

Studies that analyze healthcare administration policies at the system level fall under this category. Two main categories of research studies exist. First is evaluating and investigating the efficiency of processes, including policies, plans and procedures. Second is focusing on detection of fraud in the healthcare system. Suitable types of health data for each of these study areas, and the applicable dimensions of big data, are discussed below.

2.4.4.5.1 Process Efficiency

Any sort of systematic process within healthcare could potentially be assessed here, and improvements suggested through analysis of the associated data using data mining techniques.

Examples of this form of research include evaluation of health insurance plans by identifying significant factors involved in improving health policies [44], and identification of hospitals with potential quality problems [77]. Administrative data are considered to be the main source of information for evaluating healthcare systems. In addition, individually-directed patient-generated data could also contain useful "unofficial" information about the efficiency of various health systems, based on patients' experiences.

The collected data for this type of studies could potentially be large in scale, contain materials with different formats (e.g., unstructured text, video and structured data), and be highly uncertain in their accuracy. Optimally, health services research is conducted in a planned and orderly fashion, often repeated regularly, and hence real-time responses are probably not required.

The available data for this set of studies contains at least three dimensions of big data – volume, variety, and veracity; hence, the possibility of enhancing the quality of analysis using big data analytics is "high".

2.4.4.5.2 Fraud Detection

Claiming reimbursement for unnecessary drugs or procedures by patients, doctors or even hospitals can cause a financial burden to insurance companies and governments. Data mining techniques have aided these organizations to detect cases of fraud [54]. Administrative data contain the details of what was claimed; however, the necessity and truthfulness of the claims can be verified based on conditions and patterns existing in clinically-generated data, as well as clinically-directed patient-generated data. An instance of such fraud is a patient who has claimed the use of a drug, but based on his health conditions and background, it would be unlikely that he would need that particular drug.

A large volume of administrative and/or clinical data may be necessary for the evaluation of fraud. Ideally, inspections would be better to be done in real-time as the claim is submitted, to avoid delayed prosecution. In addition to typical uncertainties that exist in health data, such as the linkage of various data sources, there will almost certainly be contradictions which require detection.

In this sub-category, the possibility of enhancing the quality of analysis using big data analytics is "high" as the available data contains at least three dimensions of big data – volume, velocity and veracity.

2.4.4.6 Summary

Data mining is a useful analytic tool, known for its ability to be used to analyze large datasets to discover latent patterns and trends. However, traditional algorithms of data mining have limitations in handling large collections of health data. Big data analytics have been introduced to exploit these algorithms. Data mining has many different applications in healthcare, but the utility of using big data analytics for analyzing health data has not always been clear. It could be claimed that health data are a form of big data, and therefore using big data analytics is necessary when analyzing health data. However, this assumption may not be correct in all situations or with all types of health data.

The applications of data mining in healthcare were divided into four categories (clinical decision making, population and public health, genetics and biomedicine, and health administration policies), and then into multiple sub-categories. Each sub-category represents a set of studies that are common in design and requirements. The accessible types of health data (from clinical, administrative, patient-generated, and machine-generated) and the

applicable dimensions of big data (from volume, velocity, variety, and veracity) for each sub-category were presented and discussed.

In summary, not all datasets associated to every health data analysis study should be considered amenable to big data analytic techniques. Depending on the type of study and characteristics of the data, big data analytics can provide different degrees of enhancement of the quality of analysis. There cannot be one definitive statement about the efficacy of all data mining tools in healthcare because different types of health data may be used, and different analytic techniques could be exploited. Accordingly, researchers, clinicians and health policymakers should carefully examine the characteristics of the data and study conditions before making a decision on which (if any) big data analytic techniques can bring insights and add value in their study of the data.

2.5 Bias in Health Data Mining

Earlier in this chapter, we reviewed the different types of errors that can occur in every health data analysis. The systematic errors are called bias and include various types: selection bias, information bias, and confounding. The definition of each type was provided in addition to the subtypes. However, we still need to know whether these errors were handled properly in studies that apply data mining to health data. This issue was one of the main concerns which led to conducting this research study in the first place. Therefore, we ran a systematic review on the literature to fully investigate how bias is addressed in past data mining works. Our literature search plan is provided at the end of this section.

Based on our exploration in the literature, the vast majority of data mining studies which are applied to health data do not handle bias and confounding in their work. At the most, they

leave these issues to be handled at the preprocessing stage by those who extract, clean, and transform the data. The computer scientists' primary focus is therefore evaluating whether a new data mining algorithm performs well on health data or not. The findings of these studies are reported as secondary outcomes, but there are serious concerns in their validity.

Some researchers attempted to reduce the effect of existing defects in the data by manipulating the algorithms and developing new data mining algorithms that take into account the potential biases in the data. Although they did not report these defects using the same terminology used in epidemiology for the known biases, their influence and behavior are very similar. For instance, in a case where the algorithms did not consider any preferences in the associations in their design, the results contained discriminations on gender, race, health status, and other factors which could be sourced from a selection bias. Even though the investigators removed the sensitive attributes, the machine learning techniques still followed discriminative patterns due to correlated information in the input data. Therefore, a fairness-aware algorithm was designed to exclude these discriminations when mining the data [78]. On the other hand, data collected in health care often contain missing values, creating bias in the outcome models. Various SVM-based methods are proposed to build predictive models from large datasets and reduce the impact of missing values at the same time [79][80].

Data mining techniques are also employed to account for confounding issues. To investigate childhood injuries [81], one of the main reasons given for using data mining analysis is to include the numerous uncontrolled confounder variables. The lack of including all related variables is believed to undermine the validity of conclusions. In addition, there have been several attempts to develop new algorithms in the association analysis of data mining to search for and investigate confounders [82].

In a study [83], researchers at Siemens Healthcare in Portugal provided a roadmap to properly preprocess structured clinical data for predictive modelling and decision support to avoid biases. The focus was specifically on structured clinical data and prediction techniques. They were not able to find appropriate guidance in the literature in order to prepare and preprocess EHR data for data mining purposes while avoiding potential pitfalls. The main technical challenges they encountered were: (1) gathering and integrating data, (2) identifying and handling different feature types, (3) combining features to handle redundancy and granularity, (4) addressing missing data, and (5) handling multiple feature values. The strategies they considered to address these challenges included cross-checking identifiers for robust data retrieval and integration; applying clinical knowledge in identifying feature types, in addressing redundancy and granularity, and in accommodating multiple feature values; and investigating missing patterns adequately. Although the preprocessing guidelines materialize the contextual concerns on the data to technical action tasks, the gaps remaining in this valuable research study include giving guidance on other types of health data, enabling other data mining techniques, and providing directions to achieve reliable and continuous outcomes.

2.5.1 Literature Search Plan

We designed four attempts to search the literature on this issue. The initial ones were not quite successful, but eventually the last one led to more relevant articles. Six articles were selected from the results. We used PubMed, Scopus, IEEE Xplore, and Web of Science literature databases and searching tools, as the topic was related to both the fields of computer science and epidemiology/biomedicine.

List of attempts and their outcomes:

1. *(Bias and Data Mining in title/abstract)*: This returned lots of irrelevant articles due to studies just using a machine learning technique to handle a small bit of their work without being an actual data mining study.
2. *(Bias in title/abstract, Data Mining in title)*: This resulted in very limited outcomes, but some were relevant.
3. *(Bias in title, Data Mining in title/abstract)*: Again, this resulted in very limited outcomes, but a few were relevant.
4. *(Bias in title/abstract, Data Mining in title, excluding ML)*: This returned more outcomes, and some were relevant.

Title of selected articles from the results of attempt 4:

- "Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining" [78]
- "Data Mining: Childhood Injury Control and Beyond" [81]
- "Multilevel Weighted Support Vector Machine for Classification on Healthcare Data with Missing Values" [80]
- "Preprocessing structured clinical data for predictive modeling and decision support. A roadmap to tackle the challenges." [83]
- "Efficient Discovery of Confounders in Large Data Sets" [82]
- "Fast imbalanced classification of healthcare data with missing values" [79]

Queries of attempt 1 (i.e., with ML):

- ***PubMed Q1*** **=** (epidemiolog*[Title/Abstract] OR health[Title/Abstract] OR healthcare[Title/Abstract]) AND (bias[Title/Abstract] OR confound*[Title/Abstract]) AND (data min*[Title/Abstract] OR machine learn*[Title/Abstract] OR preprocess*[Title/Abstract] OR pre-process*[Title/Abstract])
- ***Scopus Q1*** **=** TITLE-ABS-KEY((epidemiolog* OR health OR healthcare) AND (bias OR confound*) AND ("data min*" OR "machine learn*" OR preprocess* OR pre-process*))
- ***IEEE Xplore Q1*** **=** (epidemiology OR health OR healthcare) AND (bias OR confound*) AND (data min* OR machine learn* OR preprocess* OR pre-process*)
- ***Web of Science Q1*** **=** TS=((epidemiolog* OR health OR healthcare) AND (bias OR confound*) AND ("data min*" OR "machine learn*" OR preprocess* OR pre-process*))

Queries of attempt 4 (i.e., without ML)

- ***PubMed Q4*** **=** (epidemiolog*[Title/Abstract] OR health[Title/Abstract] OR healthcare[Title/Abstract]) AND (bias[Title/Abstract] OR confound*[Title/Abstract]) AND (data min*[Title/Abstract] OR preprocess*[Title/Abstract] OR pre-process*[Title/Abstract])
- ***Scopus Q4*** **=** TITLE-ABS-KEY((epidemiolog* OR health OR healthcare) AND (bias OR confound*) AND ("data min*" OR preprocess* OR pre-process*))
- ***IEEE Xplore Q4*** **=** (epidemiology OR health OR healthcare) AND (bias OR confound*) AND (data min* OR preprocess* OR pre-process*)

- ***Web of Science Q4*** **=** TS=((epidemiolog* OR health OR healthcare) AND (bias OR confound*) AND ("data min*" OR preprocess* OR pre-process*))

2.6 Data Mining on Immune-Mediated Diseases

Data mining has different applications to health care, including “Clinical Decision Making” and “Population Health”, which are relevant to the motivation of this study. In the cases focusing on the prevalence of a specific disease to identify the patterns, trends, and causes of spreading that disease across a population, different risk factors and health determinants are taken into consideration, including early life, lifestyle, and socio-demographic.

We explored some of the studies which applied data mining methods or specific machine learning techniques to health data to analyze any that focused on immune-mediated disease. There were some works on asthma, few on multiple sclerosis (MS), and some on diabetes (mostly on type 2) which their findings are reported below. There was less frequent research on type 1 diabetes (T1DM), inflammatory bowel disease (IBD), and juvenile arthritis (JIA) — probably due to their extremely limited instances — which mostly included experimentation of machine learning algorithms on genetic factors in biomedicine, and not end-to-end data mining process on environmental factors.

2.6.1 Asthma

A Taiwanese study proposed a data mining mechanism for predicting sudden attacks of chronic asthma by considering both bio-signals (signals in living beings that can be continually measured and monitored) of patients, and environmental factors [84]. They constructed pattern-based decision tree and association rule algorithms and applied them to a data set of allergy patients in a hospital. They were able to build models with high accuracy. In a different

study, a patient monitoring system for asthma care using data mining was proposed and effectively predicted asthma attacks [85]. This system analyzed users' daily bio-signal records and environmental data and provided proper medical instructions and health messages. Also, there has been an attempt to identify patients whose asthma may not be well managed with the aim of improving asthma management using community pharmacy medication records in Tasmania, Australia [86]. Although said data mining applications were designed and deployed to predict the patients, mainly traditional statistical analyses were conducted on the collected data to successfully identify patients with suboptimal asthma management.

2.6.2 Multiple Sclerosis

As multiple sclerosis — like the other diseases in this category — is a multifactorial disease and the biomarkers are non-linear, traditional statistics would not be suitable and powerful enough to detect the relationships between biomarkers. Therefore, an artificial neural network method (AutoCM) was applied in one study to find the relationship between immunological markers in multiple sclerosis (MS) [87]. The outcome suggested unique immunological signatures for different MS phenotypes. Also, data mining can be applied to genetics data, such as gene expression data (microarray data) and even whole genome data (for genome-wide association studies), to identify disease-related genes. For instance, as a result of mining expressed genes in peripheral blood mononuclear cells of patients with multiple sclerosis (cases) and other neurological diseases (control), the TNFSF10 gene was significantly associated with MS. SVM methods produced classification models with the highest accuracy [88].

2.6.3 Diabetes

A systematic review of data mining technologies for diabetes included 17 published papers in MEDLINE [89]. Most of the studies were conducted on type 2 diabetes, with only a few attempts on type 1 diabetes, such as predicting blood glucose level of type 1 diabetics and analyzing genomic data [90].

Although type 2 diabetes is a non-immune disease, some data mining studies on diabetes proposed generic mechanisms with the hopes that they could be applied to both types, including analyzing diabetes-related proteins [91], identifying the best intervention in diabetes care [92], specifying study designs for clinical diabetes research in Canada [93], predicting diabetes or pre-diabetes using common risk factors in China [94] and Iran [95], and monitoring diabetes care using administrative data [96], for which the results might not be reliable for type 1 diabetes disease specifically.

2.7 Terminology

Data mining is an integration of different fields of science such as machine learning, statistics, and databases. In addition, the known concepts in epidemiology are also discussed in this research study due to its context. One of the challenges we observed in conducting this research study was that a concept may be represented with different terms, each originating from a different discipline. Using various terms is an unnecessary source of confusion for the reader. On the other hand, as a multidisciplinary study, the readers can be from different backgrounds and may not be familiar with other terms relating to the same concept.

2.7.1 Relevant Terms

The most popular terms that are relevant and sometimes used interchangeably in the context of applying data mining techniques to structured health administrative data to evaluate population health issues are listed in the following sets:

- Field, attribute, variable, dimension, feature, factor, column, property, characteristic
- Patient, individual, record, row, data point, instance, data object
- Value, data cell, piece of information
- Exposure, input variable, predictor, covariate, determinant
- Outcome, target variable, labels, classes
- Cases, cohort, diseased patients
- Controls, non-diseased patients
- Cluster, cohort, group, segment

Even though some terms in each set have a semantic relationship with others, they may refer to a slightly different concept. To clarify the usage of these terms in the focused context, we attempt to distinguish which terms in a set of relevant terms refer to a slightly different concept and which are considered synonyms. Table 2.1 summarizes the synonyms and definitions in each set of relevant terms. The synonyms can be used interchangeably.

2.7.2 Populations

To remain consistent about the types of populations throughout the study, we use the standard terms suggested in the RECORD statement [1] which include:

Table 2.1: Similarity and dissimilarity of relevant terms

Term	*Synonyms*	*Definition*
Factor	*Characteristic*	A cause, event, occurrence, or phenomenon
Variable	*Field, Attribute, Dimension, Feature, Column, Property*	The artifact referring to a factor
Individual	*Patient*	A person in the population
Record	*Row, Data point, Instance, Data object*	The series of collected data related to a person
Piece of information	-	One unit of information in the collected data
Value	*Data cell*	The number or characters that demonstrate a piece of information for analysis
Exposure	*Determinant, Predictor, Covariate*	A known factor that we tend to evaluate its influence on the concerned factor
Input variable	-	The variable representing an exposure
Outcome	-	The concerned factor
Target variable	-	The variable representing the outcome using distinct values
Label	*Class*	A distinct value used in the target variable
Cases	*Cohort, Diseased patients*	The individuals suffering with the outcome and have a true value in the target variable
Controls	*Non-diseased patient*	The individuals not suffering with the outcome and have a false value in the target variable
Cluster	*Group, Cohort, Segment*	A collection of individuals with common characteristics

- **Source population:** All individuals that fall into the population the study focuses on and intends to investigate.
- **Database population:** The individuals from the source population whose data have been collected and is available in the databases.
- **Study population:** The individuals that are finally included in the study dataset.

Figure 2.3 demonstrates the different levels of population which are used in this research study.

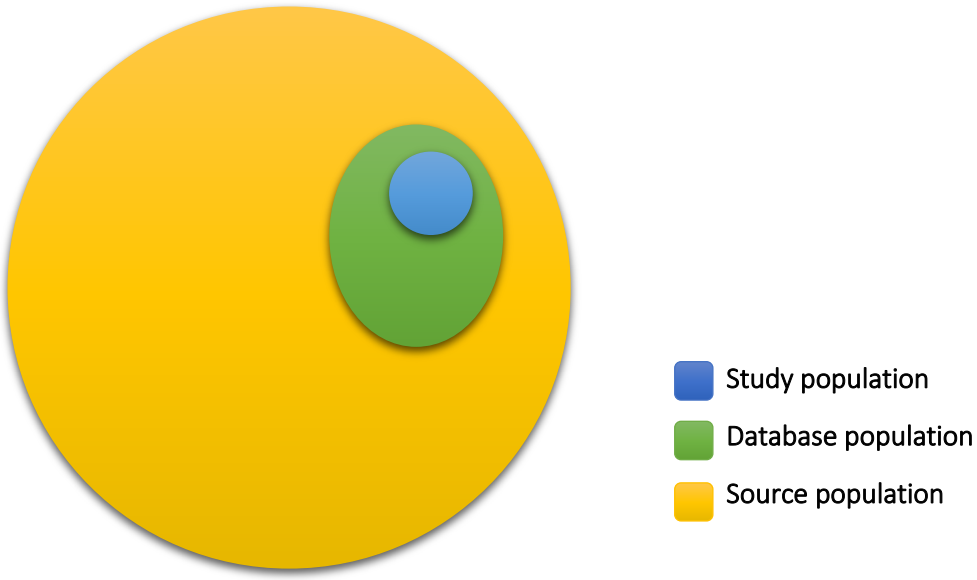

Figure 2.3: Population hierarchy in studies using routinely collected data sources [1]

2.7.3 Technical Terms

There are also some technical terms that are related to building and evaluating the data models. Again, some are used interchangeably, but they have their own specific meanings or definitions. They are clarified below to make sure they are properly used.

1. **Performance, Quality:** This concept is sometimes referred to using "accuracy" and "precision". By these terms, we mean how good the predictions are in predictive

models, and how separate and cohesive the clusters are and how strong the associations are in descriptive models.

2. **Certainty:** This refers to the performance of models considering their potential biases. If the models have high quality and low risk of bias, their certainty is high.

3. **Computational time:** This is sometimes referred to using "performance", but we differentiate between the terms. Here, we mean the time required to fit a model or the cost of building the model.

4. **Complexity:** This refers to the level of complexity of models in doing the predictions, in contrast to their simplicity. It can be identified using the structural parameters of the models. For instance, complex models include but are not limited to decision trees with lots of nodes or high depth, neural networks with lots of hidden layer nodes, and polynomial regressions with high degree.

5. **Accuracy:** This refers to a metric assessing the quality of a predictive model; there are also other metrics for the same purpose such as AUC (area under the ROC curve) or G-mean (geometric mean). It is calculated using this formula: (TP+TN)/(TP+TN+FP+FN).

    - TP: True Positive
    - TN: True Negative
    - FP: False Positive
    - FN: False Negative

6. **Sensitivity, Recall, Hit rate, True positive rate (TPR):** This means how many of the cases are accurately predicted and is calculated using this formula: TP/(TP+FN).

7. **Specificity, Selectivity, True negative rate (TNR):** This means how many of the controls are accurately predicted and is calculated using this formula: TN/(TN+FP).

8. **Precision, Positive predictive value (PPV):** This measures how many of the predicted cases are actual cases and is calculated using this formula: TP/(TP+FP).

9. **Miss rate, False negative rate (FNR):** This measures how many of the cases are predicted by mistake. It is calculated using this formula: FN/(TP+FN).

10. **Methods (in data mining):** Either predictive (supervised) or descriptive (unsupervised).

11. **Technique, Type of model:** The main techniques in each data mining method, such as classification, clustering, and association.

12. **Model family:** This refers to categories of models with the same semantic structure, such as decision trees, neural networks, hierarchical clustering, and partitioned clustering. Every model family is grouped under a specific technique.

13. **Algorithm:** This refers to a specific algorithm in each family, such as C5 and CART in decision trees.

14. **Model:** This refers to the trained model by applying an algorithm on a dataset.

15. **Final model:** This refers to the trained model using the entire training dataset, while its performance is the average accuracy of models built in various folds if the training set is resampled during the model building.

16. **Model tuning:** This refers to feeding the parameters of the algorithm with different values to find the best final model with highest average accuracy.

17. **Model building, Model training, Model fitting:** This is the process of building a model using a dataset. Similarly, “overtraining” and “overfitting” refer to the same concept, while “undertraining” and “underfitting” mean the same too.

18. **Training set, Training data:** This refers to the portion of the study dataset used for building the model.

19. **Validation set, Validation data, Temporary unseen data:** This refers to the portion of the training set that is used to estimate the quality of the model trained when a resampling method, such as cross-validation, is used.

20. **Test set, Unseen data:** This refers to the portion of the study dataset that was not used in training and was kept aside to evaluate the quality of the final model as unseen data.

3 RESEARCH METHODOLOGY

In this chapter, we comprehensively describe the plan and process used to conduct this research study. First, the problem, objectives, and research questions are reviewed. Based on the mentioned requirements, we selected the Design Science Research Methodology (DSRM) for the research study's design, which we cover next. Finally, we clarify the scope of this research and introduce the outcome of the study.

3.1 Problem

As mentioned in section 1.1, our motivation to conduct this research study started with the alarming rise in the prevalence of immune-mediated inflammatory diseases (IMIDs) [3]. Many of these disorders are of higher prevalence in Westernized nations such as Canada, with low prevalence noted in people living in developing nations [4][5][6]. Exposure to some environmental factors early in life has been associated with increased risk of these diseases. However, as noted in the introduction, hypothesis-driven analyses may not identify all risk factors, and may not adequately explain variable interactions or time-related effect modification of variables on the risk of disease.

While data mining has the potential to change the way health care problems are approached [11], the techniques are still not popular nor trusted among epidemiologists. After an initial investigation of how typical epidemiological studies are conducted compared to how data mining techniques are applied to health data in computer science papers, it was not difficult to understand why data mining methods are not easily trusted by health domain experts. A

research study was required to further explore the types of epidemiological studies, the types of health data, the sources of errors accounted for in epidemiological studies, the analytical methods of data mining, and the typical process of conducting a data mining study. The study would result in designing an integrated solution to provide reliable results when applying data mining techniques to health data.

Of course, the domain of health data usage and epidemiological study topics is broad. At the same time, data mining methods are more suitable in some particular epidemiological study designs than others. Chapter 2 fully discusses and elaborates on those topics as background knowledge for the thesis. Therefore, we narrowed the scope of our investigation to be able to focus on a measurable problem that fits into a PhD research program. Briefly, the scope was set to apply data mining techniques on health administrative data, whether population-based or not. The potential areas and details of our scope are discussed further in section 3.4.

In summary, the core problem this study is focused on is the reliability of data mining on health data. There are numerous publications that have applied data mining techniques to different types of health data. However, it seems there are serious concerns about the validity and usability of results in many of these works. This is disappointing as data mining brings the benefit of analyzing large datasets with many variables to find patterns to address health issues, which are often multifactorial. If there is a solution for this trust problem, it will open the door to an enormous number of data analysis studies applying various data mining techniques to health data for different specific purposes. Therefore, such a solution would enable scientists to benefit from the great advantages of data mining over traditional statistics, including hypothesis-generating study and analyzing large amounts of data. In other words, the findings resulting from implementing this solution can lead to generating new

epidemiological hypotheses which can be separately examined, and to building valid models to predict the risk of new patients or specific cohorts.

3.2 Objectives and Research Questions

This research study was initiated to address the problem described in section 3.1. The main objective is to design a solution to reliably analyze health administrative data using data mining techniques. There are contextual considerations that are included in every epidemiological study in order to provide valid and reliable results. Normally, these considerations are not taken into account in studies that apply data mining to health data. Therefore, the main objective is to prepare a reliable solution — from the perspectives of both a data miner and an epidemiologist — to mine health administrative data and provide valid and trustworthy findings. To achieve this objective, the following research questions are defined:

- What are the missing parts in a typical data mining study that are accounted for in many epidemiological studies in order to provide trustworthy results?
- Can we design a reliable data mining methodology to produce valid findings when analyzing health administrative data while addressing the related epidemiological considerations?

Furthermore, we were motivated to shed light on the investigations around immune-mediated diseases when we commenced this research study. So, we ensure the designed solution is pragmatic and validate it by implementing it on real-world data as case studies. In other words, the second objective is to conduct experiments that will apply the designed solution on real-world health issues using data from children diagnosed with immune-mediated chronic diseases. The root causes of these disorders are unknown. We will focus on pediatric patients

as our target population as they are exposed to fewer environmental factors, thereby reducing the complexity of the relationships between variables and increasing the chance of making new discoveries linking environmental risk factors and disease risk. For the second objective, the following research questions were considered:

- How can this methodology be applied to health administrative data to generate new hypotheses on the association between IMIDs and early life and environmental factors?
- What are the significant early life and environmental factors that are associated with increased risk of IMIDs using data mining techniques?

3.3 Research Design

Now that we have clearly defined this study's objectives and research questions, the next step is creating a research design to plan for and conduct the study. Based on the features expressed in the objectives, the expected outcome, and the required approach to properly address the research questions, we selected the Design Science Research Methodology (DSRM) as the research methodology of this study. DSRM is an outcome-based research methodology within the field of information systems and technology. It focuses on development of an innovative artifact, such as a design methodology, with the significant intention of improving its performance based on a set of pragmatic research objectives [97], [98]. The main goal of design science research is to generate knowledge that experts of a field can use to build solutions for their problems, as opposed to explanatory sciences that develop knowledge that describes and predicts [98]. This approach is exactly what we intend to accomplish in this research study as we build an innovative method that can be used to build solutions (e.g., data mining models) to reliably address problems in the health care domain.

Hevner et al. [17] defined seven guidelines for a design science research. For each, we quote their brief definition below and describe how our study relates to it.

3.3.1 Design as an Artifact

"Design-science research must produce a viable artifact in the form of a construct, a model, a method, or an instantiation." [17] The methodology we designed in this research study is the "viable artifact". It is a new approach in conducting analysis on health administrative data using data mining techniques. Therefore, it can be applied to various case studies as separate experiments.

3.3.2 Problem Relevance

"The objective of design-science research is to develop technology-based solutions to important and relevant business problems." [17] Health care problems are among the vital issues human beings are involved with in their lives. In addition, health systems in both public and private sectors tend to operate efficiently and effectively day-to-day using new findings and discoveries. The reliable data mining methodology built in this research study is a "technology-based solution" and a step toward discovering new hypotheses and findings to assist in solving health care-related problems. For instance, if a reliable data mining approach can identify environmental factors leading to IMIDs, it may result in interventions to reduce the risk of IMIDs in children.

3.3.3 Design Evaluation

"The utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods." [17] The designed methodology employs known concepts from the data mining and epidemiology fields with demonstrated "quality and efficacy" in

conducting data analysis studies. However, it provides a new approach in using those concepts via the defined process stages and newly crafted guidelines. Therefore, the designed methodology is applied to real-world data so that its steps can be validated and improved.

3.3.4 Research Contributions

"Effective design-science research must provide clear and verifiable contributions in the areas of the design artifact, design foundations, and/or design methodologies." [17] A real gap exists in using data mining methods in epidemiological studies, which is the core venue for studying population health issues by analyzing collected health data. We believe this gap exists due to concerns about the black box nature of data mining techniques and the inability to easily trust the outcomes. The designed methodology is a new approach to conducting data mining studies reliably on health administrative data. Its application to real-world data is the verification to demonstrate it works as designed to provide reliable results.

3.3.5 Research Rigor

"Design-science research relies upon the application of rigorous methods in both the construction and evaluation of the design artifact." [17] The methodology designed in this research study has roots in well-known methods such as CRISP-DM and epidemiological data investigations. It combines these methods and customizes them based on the requirements. Without in-depth research on the advantages and disadvantages of the related methods in both fields, it was impossible to integrate the methods and provide a new solution.

3.3.6 Design as a Search Process

"The search for an effective artifact requires utilizing available means to reach desired ends while satisfying laws in the problem environment." [17] There is no better way to describe the

bottom line of the designed methodology than this quote. The designed methodology is a great example of leveraging the "available means" — data mining techniques — to enable analyzing large datasets with many variables related to a multifactorial health issue. At the same time, the core attempt of the designed methodology is to comply with the epidemiological considerations to satisfy the requirements of providing reliable results in the field of epidemiology.

3.3.7 Communication of Research

"Design-science research must be presented effectively both to technology-oriented as well as management-oriented audiences." [17] The context of this study is health care, not management. However, the designed solution is an integration between a technical and a contextual field of science. Therefore, there are two separate groups as the audience of this study — one with computer science backgrounds and the other with public health backgrounds. Even this research study's advisory committee members relate to one or the other of these fields and each had their own specific concerns. Therefore, the designed methodology and its sample applications had to be presented to two groups of audiences and comply with both sets of technical and contextual expectations.

3.4 Scope

Data mining has been applied to health care for numerous objectives. Two recent review studies that aggregate and discuss these applications are [35][38]. We have also searched through related works and compiled the applications of data mining to health care in a book chapter [18] to investigate whether using big data analytics provides serious benefits in these applications. We categorized applications into four groups based on the area of health care to which they have been applied: Clinical Decision Making, Population and Public Health,

Genetics and Biomedicine, and Health Administration Policies; these are presented in more detail in section 2.4.1. Each group was categorized into subgroups based on the characteristics of the analysis, such as requirements, scalability, and variety of data. In addition, we discussed different types of health data — reviewed in section 2.1.2 — and which of them are accessible in each of the subgroups. And finally, we discussed which dimensions of big data are applicable to the subgroups in order to determine the level of benefits gained in implementing each type of application on a big data analytics environment. Figure 2.2 summarizes the discussion and findings mentioned in that section. That review helps better define the scope of this study.

In this PhD research study, we intend to leverage the available data mining techniques and design a reliable data mining methodology for health administrative data and apply it on real-world data. Therefore, the scope of this study is to use traditional data mining techniques — and not any other analytical tools, such as big data analytics — on health administrative data — and not other types of health data, such as clinical data — for any type of application possible. In other words, only the characteristics of data mining techniques and health administrative data are taken into consideration when customizing the methods and guidelines in this methodology. To be more precise, we have broken down the aspects included in this research study's scope and defined them as follows:

- **Health administrative data:** The data that are often collected routinely as a matter of running health care systems, such as billing and administration, or non-routinely, such as registries. The health administrative data might be population-based (which includes records of all eligible individuals in a geographical jurisdiction, such as a country,

province, city, or town, or even a certain denominator of a particular demographic, like women) or not population based.

- **Data mining:** The analytical techniques that integrate machine learning, statistics, visualization, artificial intelligence, and database technologies to discover hidden patterns in structured data.
- **Reliable methodology:** A system of methods that provides trustworthy results and findings, regardless of whether the findings are accurate and useful, or low quality and useless.

Furthermore, as mentioned in section 1.1, the initial motivation for conducting this research study and designing a new data mining methodology was the increased prevalence of immune-mediated disease patients. Analyzing the data related to this family of diseases falls in the second category (Population and Public Health) of data mining applications for health care. More specifically, the case studies that we implement the designed methodology in belong to the Developed Conditions and Limited Conditions study groups of the mentioned category. Predicting asthmatic patients requires collecting data from the patients as the disease develops, and, at the same time, there are frequent cases with this disorder in our target population (i.e., Ontario, Canada); therefore, it has the same characteristics as the studies that are grouped in the Developed Conditions subcategory. On the other hand, predicting pediatric IBD in young children falls in the Limited Conditions bucket of studies as there are a small number of cases in the target population. However, we must still allow time for the condition to develop and to collect the related data. Figure 3.1 displays the scope of the case studies we have planned to address, implementing the designed methodology on their related data and running experiments.

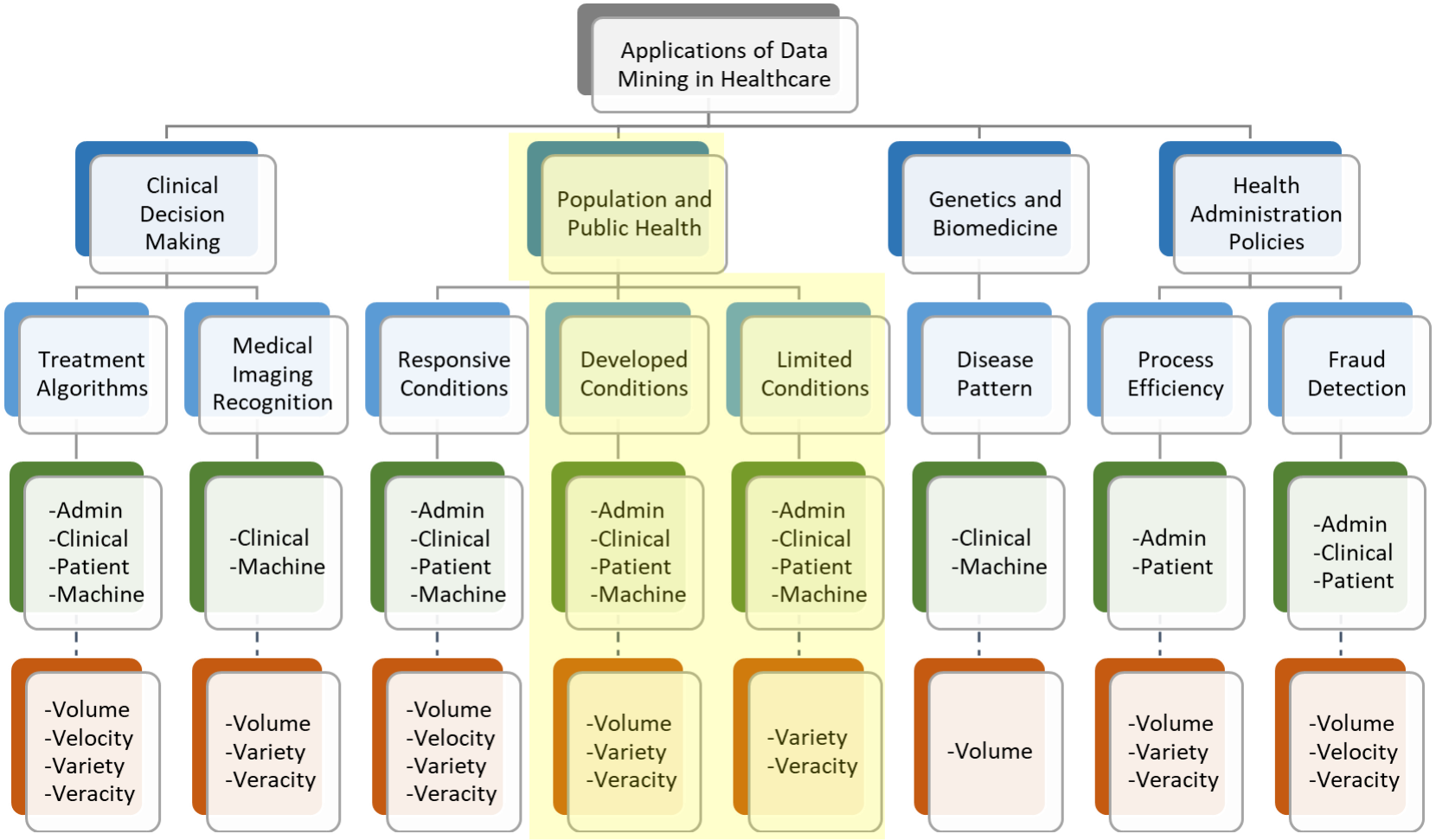

Figure 3.1: Scope of case studies (highlighted in yellow) intended for implementation of the designed methodology in this research using the hierarchy from our book chapter [18]

3.5 Solution

To achieve the objectives of this research study, we designed a generic, quantitative data mining methodology that can be implemented to reliably analyze health administrative data. Throughout this thesis, we refer to this methodology as the reliable data mining methodology. This methodology is mainly derived from the CRoss Industry Standard Process for Data Mining, aka the CRISP-DM process model, introduced in section 3.5.1. We simplified the phases of CRISP-DM into four practical stages of a data mining process. Furthermore, we added a context layer to customize the process for analyzing health administrative data by considering additional components that provide reliability in the results. The integration of a derived version of CRISP-DM and the context layer generated the reliable data mining methodology.

The main characteristic of this methodology is to reliably execute the analysis process and provide reliable results. It is important to keep in mind that reliability does not necessarily mean high accuracy models. It means being able to trust the final conclusions — whether they are claiming to have valid results or invalid results (including inaccurate models). In a nutshell, this attribute is achieved by integrating epidemiological considerations to avoid sources of error and comply with domain validations, in addition to running the technical data mining steps carefully and precisely to provide valid and fair models — even if their accuracy is not high. These items have been crafted into the designed methodology, especially in the preprocessing guidelines.

The reliable data mining methodology is fully presented in chapter 4. However, before proceeding with the designed methodology, we review the CRISP-DM process, the advantages of this methodology, and the challenges in designing this solution.

3.5.1 CRISP-DM

CRoss Industry Standard Process for Data Mining (CRISP-DM) was initially introduced as process lifecycle for data mining in 2000 [99] and is still the most well-known and widely used data mining process model in both academic papers and industry reports [100], [101]. One of the main reasons for its success is that it is industry, application, and tool neutral [101].

CRISP-DM breaks down a data mining process into six major phases as demonstrated in Figure 3.2. Here, we briefly present the requirements and actions in each of these phases.

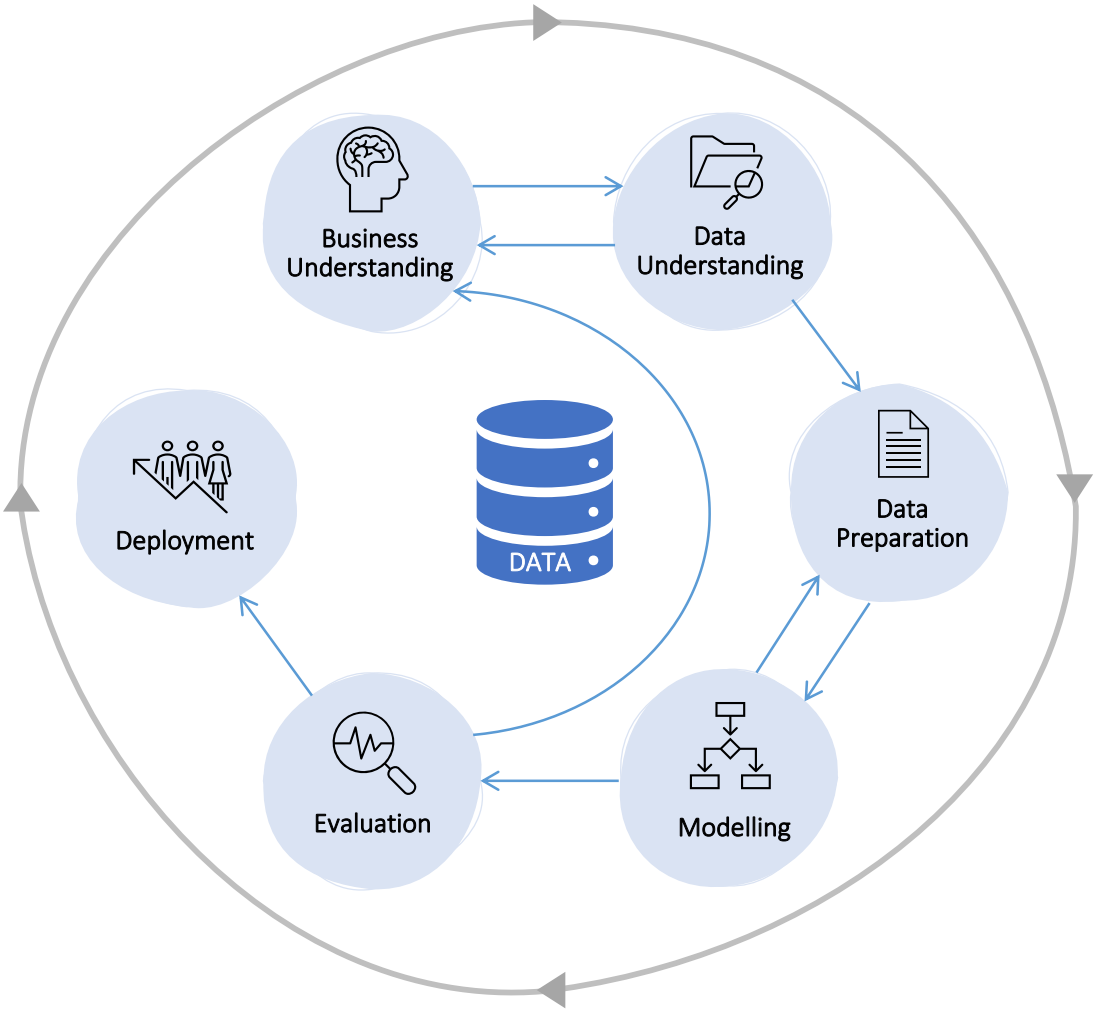

Figure 3.2: Six major phases of the CRISP-DM process model [99]

- **Business Understanding:** The project objectives and requirements from a business perspective are understood in the first step. Then, those are converted into a data mining problem definition, and a plan is designed to achieve the objectives.
- **Data Understanding:** The second phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, identify data quality problems, discover first insights in the data, and detect interesting subsets to form hypotheses for hidden information.
- **Data Preparation:** In this phase, all activities to construct the final dataset from the initial raw data are covered. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data.
- **Modelling:** In the modelling phase, various modelling techniques are selected and applied, and their parameters are calibrated to optimal values. There are typically several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data, which often require stepping back to the data preparation phase.
- **Evaluation:** At this stage, a high-quality model from a data analysis perspective has been built. Before proceeding to final deployment of the model, it is important to thoroughly evaluate the model and review the steps executed in its construction in order to make sure it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

- **Deployment:** Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In most cases, the customer is the one who will carry out the deployment steps. Even if the analyst deploys the model, it is important for the customer to understand up front the actions which will need to be carried out in order to actually make use of the created models.

3.5.2 Advantages

The reliable data mining methodology attempts to include and incorporate the advantages of both its technical and contextual fields. These benefits are listed as follows.

- **Data Mining**
    - Analyze the entire population as opposed to a sample: Data mining methods have lower computational costs compared to pure statistical methods and perform faster due to their heuristic functions and fuzziness in results. Therefore, data mining allows including more records in building the models, increases the chance of finding patterns from a more complete set of individuals, and requires fewer sampling attempts.
    - Include many variables in modelling: As most health issues are multifactorial and complex, this ability significantly helps to discover more contributing factors that have impact together. In addition, it reduces the risk of confounding.
    - Extensive modelling algorithms: Data mining has developed rapidly in the past few decades. As a result, a broad range of algorithms from various involved fields are available for different analytical purposes.

- Hypothesis-generating data-driven analysis: These methods are an excellent means of generating new hypotheses that can be tested separately.

- **Epidemiological Studies**
  - Use validated algorithms to identify cases more accurately.
  - Account for sources of bias when designing the study.
  - Include individuals based on verified criteria to investigate the population properly, instead of incautiously analyzing any records available in the collected data.
  - Verify the validity of results by evaluating the errors and biases that exist in the study and data, in addition to running statistical tests of significance and examining p-values.

3.5.3 Challenges

As described earlier, this research is an interdisciplinary study which aims to design an integrated solution. The two fields of science involved in this study come from quite different backgrounds. Data mining, even though a multidisciplinary field, is driven by computer scientists with mathematical and engineering backgrounds. On the other hand, clinicians and epidemiologists have a biological and medical background. Building a common ground so that constructive communication can happen between the two fields was the main challenge in conducting this research. Each side of the parties used different terminologies and analytical methods, and even had different expectations. Therefore, we had to spend good amount of time with the experts from both sides to better understand what they mean and what they focus on. Sometimes, basic scientific concepts were barriers to digesting their common approaches. Therefore, designing a solution that can satisfy each side of the audience was not easy.

However, both sides were interested and passionate to see whether an appropriate link could be created so that each side could benefit from it. Overall, communication and satisfaction were the main elements of challenge in conducting this research study.

4 THE RELIABLE DATA MINING METHODOLOGY ON HEALTH ADMINISTRATIVE DATA

In this chapter, we discuss the fundamentals of a new reliable data mining methodology and the steps taken to design it. As the proposed methodology is intended to be a generic solution, we define the boundaries within which it is expected to operate. Finally, we go through the details of all the methodology's defined stages.

4.1 Boundaries

Though very attractive to analyze by data miners and computer scientists, health data are complex. In such circumstances, data mining research papers that apply a new algorithm to a health dataset with the purpose of testing that algorithm often claim new health findings as results without validating them. The lack of validation could be one of the main reasons that epidemiologists are hesitant about the performance of data mining techniques in general. Therefore, it is important to first understand the different types of health data and consider their characteristics, which are discussed thoroughly in section 2.1. In this research study, we focus on health administrative data as there are enormous amounts of this type of health data being generated every day as a matter of running digital health systems. Health administrative data are not collected to answer any predefined research questions and do not contain detailed clinical information; however, they can still be a great source of information as a secondary use, especially when they are population-based and cover all varieties within a particular source population.

In addition to the specific type of health data, it is important to address how data mining techniques fit into epidemiological studies as the related topics in this scope are typically addressed in the field of epidemiology. Epidemiology is a science, and it is naïve to claim that we attempt to enable data mining techniques for all different types of studies in this field. In addition, it is important to understand how each type of epidemiological study is designed, which we present in section 2.2, so we can compare them with typical data mining methodologies. Again, this helps us to better carve out the scope based on the types of epidemiological studies that our solution is compatible with. Data mining techniques might not even be applicable to all types of epidemiological studies and each type might require a different set and sequence of actions — that is, a different methodology.

We need to consider that data mining is also a multidisciplinary field of science with a broad variety of techniques, each originated from a different discipline. Therefore, we need to make sure which type of data mining models are possible to be built to accommodate the considerations from epidemiological studies. Furthermore, similar to how incautiously different types of health data are used to verify new algorithms in data mining studies, the range of applications of data mining to health care is also usually underestimated. It might be believed that all analytical tools can be employed, no matter what the type of application is. We reviewed the applications of data mining to health care in section 2.4.1 and elaborated them so that their characteristics are well understood, in addition to which types of health data are relevant to each and whether big data analytics can play a significant role. This helps us to set our expectations wisely from the proposed solution.

These topics were the concerns in regard to the boundaries that we tried to consider when designing the methodology which is discussed in the next section.

4.2 Design

A derived version of CRoss Industry Standard Process for Data Mining (CRISP-DM) is presented in Figure 4.1, wherein there are four main steps: data selection, preprocessing, modelling, and evaluation. In the first step, the datasets and attributes are selected. Then, in the second step, the data are preprocessed following a set of instructions. The required models are built in the third step, and finally, they are evaluated. Although a Business Understanding step, which aims to define the objectives, exists in the CRISP-DM process, we require an additional layer to this stream that addresses the specific concerns applicable to studies relevant to the defined scope. We call this the context layer which initiates from the problem and contains a contextual preprocessing in addition to the known technical preprocessing. We also added a findings evaluation in addition to the regular model evaluation.

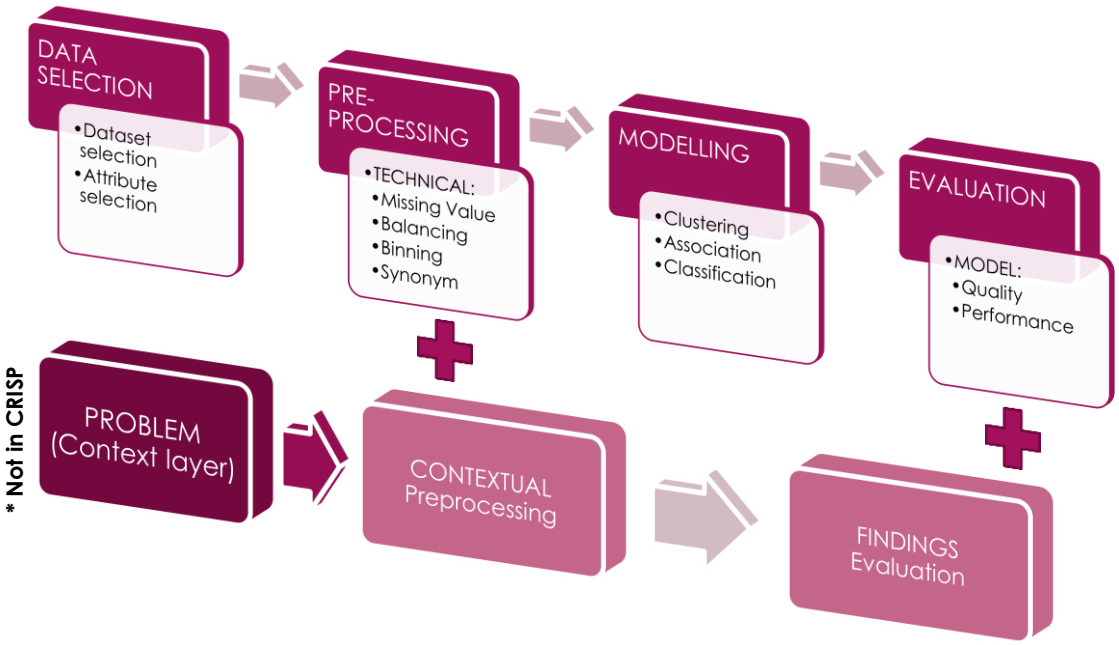

Figure 4.1: Context layer added to CRISP-DM

The context layer mainly aims to enhance the CRISP-DM process to address its gaps. Based on our investigations, there are three characteristics in past data mining works on health data that do not provide enough confidence to domain experts to gain their trust: impartiality, validity, and sustainability. Therefore, we attempt to design a new methodology that contains these three key features to provide reliable results.

The proposed reliable methodology, as the integration of the CRISP-DM process with the context layer and the relations between inner components, is displayed in Figure 4.2, with the new components in black. These new components represent the impartiality, validity, and sustainability features we plan to embed in this methodology. Each stage is described below.

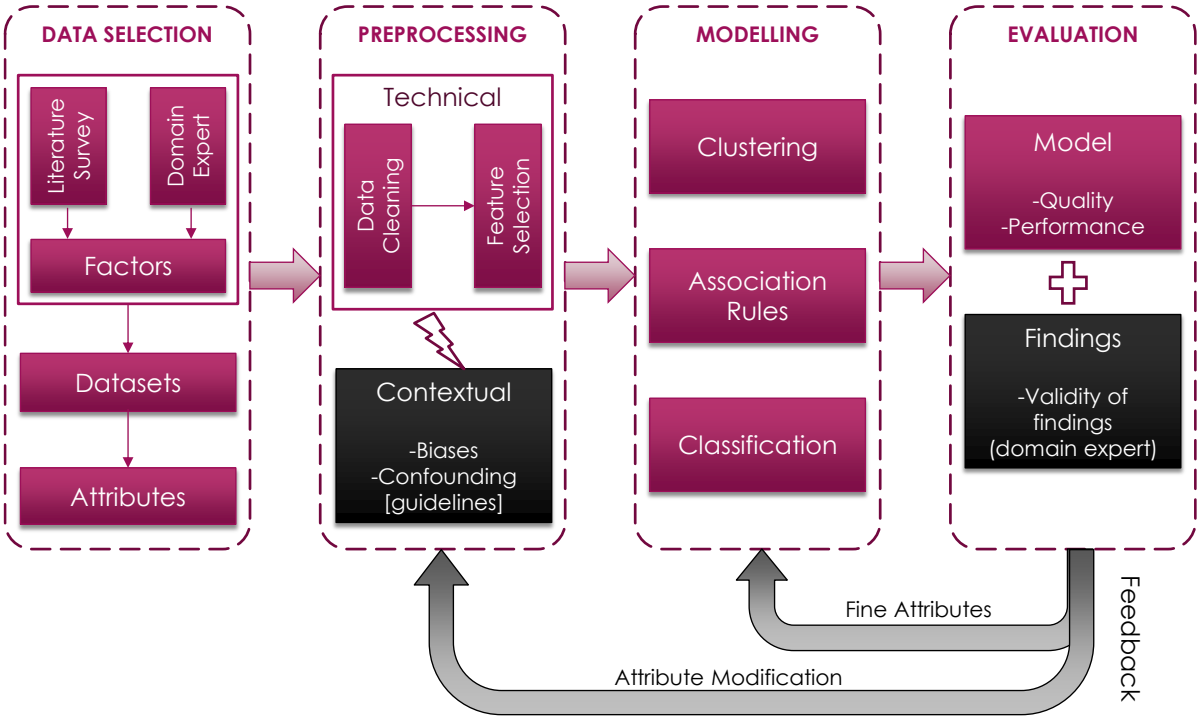

Figure 4.2: Stages of the reliable data mining methodology

4.2.1 Impartiality

Based on our understanding from the literature review described in section 2.5, data mining techniques are incautiously applied to health data after some technical massaging of the data just to satisfy the algorithm requirements in vast majority of data mining studies. But not enough attention is paid to which individuals are included or missed in this data. Furthermore, any type of real-world data, including health data, usually includes various types of random and systematic errors that are quite impossible to detect out of context, in addition to the obvious noise that is handled as part of the technical preprocessing. Therefore, the designed methodology needs to be capable of conducting an unbiased data mining study.

4.2.2 Validity

Health issues are usually very complex to analyze and require considering different aspects relevant to the issue such as previous hypotheses, genetic causes, and environmental effects. Therefore, reviewing the results with the domain expert and verifying the findings under different conditions are both necessary to make sure the results are valid.

4.2.3 Sustainability

As has been embedded in the design of the CRISP-DM process, data mining can be utilized in a continuous fashion. Data mining papers that focus on designing, developing, and evaluating particular algorithms to build predictive or descriptive models might apply their techniques to the health data to find the desired information and then stop the process. However, we retain the continuous process embedded in the CRISP-DM in this methodology by customizing it to analyzing health administrative data in order to keep the process sustainable.

4.3 Process

In this section, we go through the details of the designed methodology as a generic solution to reliably apply data mining techniques to health administrative data and provide valid results. Note that valid results do not necessarily mean models with high accuracy or useful results; rather, it means you can rely on the results as the risk of error has been reduced by eliminating the potential sources or at least reporting them along with the results, regardless of whether the model results are acceptable or not. Therefore, the reader can trust the results and make appropriate decisions based on them.

The preprocessing guidelines designed in Stage 2 of this methodology are discussed in the next chapter as they are a significant part of this methodology and one of the key contributions of this thesis which need to be presented in detail. In addition, the new components designed in this methodology (demonstrated with black elements in Figure 4.2) attempt to provide the key features — impartiality, validity, and sustainability — which are respectively discussed in Stages 2, 4, and 5.

4.3.1 Stage 1: Data Selection

At the first stage, the analyst needs to select the required data to address the concerned problem or research questions. There are three major tasks at this stage: 1) selecting the relevant factors, 2) selecting datasets from source databases, 3) selecting available attributes. We need to keep in mind that health administrative data are already collected, and we use them for secondary purposes retrospectively.

It is recommended to first compile a list of factors relevant to the outcome or target based on related works in the literature, in addition to opinions of domain experts. In cases where the

concerned problem is related to a health condition such as a disease, the list of risk factors and protective factors can be gathered. Note that each factor could be represented by one or many attributes, depending on the availability of data in every study.

To continue, the required data are chosen from the previously collected datasets in accessible source databases. As the data are not gathered based on the concerned problem or research questions, it is likely that many of the requested factors do not have any representing attribute in the selected datasets. Therefore, a round of attribute selection must happen in order to identify which attributes in the selected datasets match the requested factors. Note that it is possible that a combination of attributes or a derived version of them available in the selected datasets could represent a factor of interest. In these cases, all these attributes are selected at this stage and are handled in the preprocessing stage. Finally, we checkmark the factors that are available in some form in the selected datasets, and we also keep a record of the ones not available as it will be needed in the next stages as part of the reliability of the methodology in providing valid results. All the aggregated information in this stage can be the documented and presented in a data creation plan.

The data miner would need assistance from a domain expert to select high quality articles from the literature, a database administrator to learn about the available datasets and their relationships, and even an analyst who is more familiar with the databases to better understand what information is stored in each attribute and how they can be interpreted.

4.3.2 Stage 2: Preprocessing

After the required data are selected, it needs to go through a series of preprocessing actions to derive a cleaned dataset that is appropriate for modelling. In preprocessing, we mainly deal with the quality of data by addressing these questions:

1. What issues in the data should we worry about?
2. How can we detect these issues within the data?
3. What can we do to address these issues?

Data preprocessing is typically a tedious process in all data mining applications. However, it becomes even more complex, sensitive, and time-consuming in health care due to the nature of health data. There are numerous attributes collected from various sources in different formats, in addition to remarkable privacy protection concerns that need to be addressed to access real-world datasets. Lastly, the findings are highly affected by the quality of data and impact the health of people which has significant consequences.

The preprocessing steps that are common in most data mining processes are mainly technical. In other words, the data are only preprocessed and prepared to become compatible with the requirements of the modelling algorithms. In this methodology, we have developed the Technical Preprocessing Guideline to define the required technical preprocessing steps relevant to health administrative data in a standard and chronological order. The technical preprocessing actions can be divided into two main categories: data cleaning and feature selection. Tasks such as dealing with synonym columns, correlated attributes, significant missing data in columns, minor missing data in rows, binning continuous fields (if needed), and balancing the data based on the target variable for classification models are applied for data cleaning purposes. It is also possible that the analyst must run some feature selection techniques to reduce the number of attributes, such as removing attributes using deep learning and genetic algorithms, replacing attributes using principal component analysis (PCA) calculations, or even deriving new attributes from the selected ones. Table 4.1 summarizes the

chronological steps defined in the Technical Preprocessing Guideline (TPG), including the cause, suggested actions, and the type of modelling or dataset that each step applies to.

As discussed earlier, there are serious concerns about the inclusivity and quality of the selected data that can cause errors in the analysis and results. For instance, it is critically important to determine which individuals are included or excluded in the prepared dataset, how they were selected, and what risks could be faced by doing so. Therefore, the systematic errors that could be included in the data need to be identified and addressed as much as possible. These potential systematic errors introduce various sources of bias and confounding to the study that are presented briefly in section 2.2.2.

We developed the Contextual Preprocessing Guideline in this methodology to inform the analyst about the potential sources of biases and confounding specifically in applying data mining techniques to health administrative data, and we provide a set of suggestions on how to measure and handle them properly. This guideline makes sure the best decisions have been made in preparing the dataset, and if there are any opportunities to mitigate the risk, those technical steps get redone. We strongly believe that if this section of the methodology is implemented properly with considerable amount of attention and effort, it makes the process of evaluating the validity of results less intensive, which then becomes just a final check on the findings. Table 4.2 summarizes the steps defined in the Contextual Preprocessing Guideline (CPG), including the source of bias that each step attempts to identify and mitigate.

Both the Technical Preprocessing Guideline and the Contextual Preprocessing Guideline are presented in full detail in the next chapter.

Table 4.1. Summary of the Technical Preprocessing Guideline (TPG)

Step	*Title*	*Applies to*	*Cause*	*Suggested Actions*
1	Dataset Creation	✓ Predictive ✓ Descriptive	Health administrative data are collected in various structures and formats that might not be suitable and understandable by the data mining techniques.	1. Select the base dataset 2. Transform into a tabular dataset 3. Apply inclusion/exclusion criteria
2	Feature Construction	✓ Predictive ✓ Descriptive	Health administrative data contain only necessary and particular pieces of information captured as a matter of running a health administration system.	Generate all required variables that represent the selected factors
3	Exploration and Planning	✓ Predictive ✓ Descriptive	Variables and records need to be investigated and plan on how to properly preprocess them.	➢ Form a planning spreadsheet and define the types, roles, and actions ➢ Eliminate the intermediary variables
4	Data Partitioning	✓ Predictive	To properly evaluate the performance of models, the data should be partitioned into training and test sets as soon as the initial structure of the dataset is prepared.	Randomly split the data into training and test sets with a 70/30 percentage proportion, or 80/20 if the dataset is not large
5	Data Cleaning	✓ Predictive (training*/test) ✓ Descriptive	Data mining techniques expect the data to be well-prepared and consistent. Health administrative data typically contain noise, inconsistency, and various formats as they are real-world data.	➢ Fix case sensitivity issues ➢ Harmonize blank cells ➢ Fix data format issues ➢ Remove invalid outliers ➢ Fix noises
6	Missing Values	✓ Predictive (training*/test) ✓ Descriptive	Most data mining techniques cannot handle missing values. Health administrative data are gathered from multiple sources that could contain incomplete information shown as missing values.	*Input Variables* 1. Keep if the model can handle them 2. Replace meaningful missing values 3. Remove variables with high missingness 4. Truncate records with high missingness 5. Impute the rest and verify (maybe after step 9) *Target Variables* o If target is constructed, fix the criteria o If not, truncate record unless it can be replaced with a valid value

7	**Categorical Variables**	✓ Predictive (training*/test) ✓ Descriptive	Depending on how health administrative data are collected, it is likely to have categorical variables. All data mining algorithms are based on mathematical equations; therefore, all values should be in numeric format to run the calculations.	➢ Transform all variables to numeric format ➢ Merge conceptually related variables ➢ Apply binning in variables with many distinct values, if needed ➢ One-hot encode all categorical variables
8	**Feature Reduction**	✓ Predictive (training*/test) ✓ Descriptive	As the collected health administrative data source from various origins, the data could have irrelevant to our research or duplicate variables which would confuse the models in finding patterns.	➢ Remove derived variables ➢ Remove unary variables ➢ Remove almost-unary variables, if possible ➢ Remove unique variables ➢ Remove highly correlated variables
9	**Data Normalization (Feature Scaling)**	✓ Predictive (training*/test) ✓ Descriptive	Every variable has a different range of values and the variables with large numbers would be dominant.	o Rescaling (min-max normalization) o Standardization (z-score normalization)
10	**Data Balancing**	✓ Predictive (training)	Classifiers should be fed with balanced datasets.	o Undersampling o Oversampling o Matched case-control o SMOTE
11	**Resampling Data**	✓ Predictive (training)	A segment of the training set is considered as the validation set (aka temporary unseen data) to estimate the model's performance.	Use 10-fold cross validation
12	**Runtime Filtering**	✓ Predictive (training*/test) ✓ Descriptive	Need to filter out some variables to investigate their impacts on the models.	Exclude unrequired input variables from modelling

** Actions are planned only based on the training set, then applied to both training and test sets*

Table 4.2. Summary of risks addressed in the Contextual Preprocessing Guideline (CPG)

Step	*Title*	*Risks*
1	Selected Source Databases	❖ *Selection bias:* Ascertainment bias ❖ *Information bias:* Misclassification bias
2	Criteria of Study Population	❖ *Selection bias:* Inclusion and exclusion biases, matching bias
3	Diagnostic Algorithms and Data Completeness	❖ *Selection bias:* Diagnostic bias, prevalence-incidence bias ❖ *Information bias:* Diagnostic bias, prevalence-incidence bias
4	Use of Diagnostic and Procedural Codes	❖ *Information bias:* Misclassification bias
5	Time-Varying Variables	❖ *Selection bias:* Survivor treatment selection bias, diagnostic bias ❖ *Information bias:* Diagnostic bias, immortal time bias
6	Lack of Contributing Variables	❖ *Confounding:* Unmeasured, measured
7	Handled Missing Values	❖ *Selection bias:* Exclusion bias ❖ *Confounding:* Measured
8	Partitioned, Balanced, and Resampled Data	❖ *Selection bias:* Non-random sampling bias, matching bias
9	Model Results	❖ *Selection bias:* Publication bias

4.3.3 Stage 3: Modelling

Normally, and for most studies in the scope this methodology is being designed for, the dataset is created and then prepared in order to build various descriptive and predictive models. Therefore, it is likely that both descriptive and predictive models are built in a meaningful sequence in the same project. Depending on the objective of the project, different data mining techniques and algorithms are employed to build these models. Every data mining algorithm has certain abilities. There is no one algorithm that works the best on all types of data or applications. The analyst usually chooses some initial algorithms, then compares the quality

of the models to identify which algorithms performed better. This is known as the "no free lunch" rule. Some recommended data mining algorithms are listed in Table 4.3. Note that this table only includes single-label classifiers.

Table 4.3: Recommended data mining algorithms in analyzing health administrative data

Method	Technique	Model Family	Algorithm
Predictive	Classification	Decision Trees	• C5.0 • Classification and Regression Tree (CART)
		Neural Networks	• NNet • Multilayer Perceptron (MLP)
		Support Vector Machines	• SVMpoly
		Regressions	• Logistic
		Other	• Random Forest • Naïve Bayes • k-Nearest Neighbors
	Estimation	Regressions	• Linear • Multivariate
Descriptive	Clustering	Partitioning	• k-Means
		Hierarchical	• Agglomerative
		Density-Based	• DBSCAN
	Association	-	• Apriori
	Outlier/Anomaly Detection	-	• Local Outlier Factor (LOF)

When a data mining algorithm is applied to a training dataset, a data mining model is built. To build predictive models, the dataset should be partitioned into training and test sets. As discussed in section 5.1.4 in the Technical Preprocessing Guideline, the technical preprocessing actions should first be applied to the training set, and then to the test set in the same sequence but based on the preprocessing parameters and models from the training set. In addition, the tuning parameters for each algorithm can receive various values so that the

model is tuned during modelling while being cross validated to produce the best final model possible. In order to measure and compare the quality of models during cross-validation and model tuning, we suggest using a metric that is sensitive to both sensitivity and specificity, such as AUC or G-mean. Finally, make sure the positive class in the target variable is specified. In some environments, the first value is considered as the positive or concerned class. This will help to get the accuracy of the concerned class in the "sensitivity" calculation.

Attribute ranking can be done either dependent or independent of predictive models to identify the important variables using various techniques. If multiple predictive models are built and the attributes are ranked using a calculation (such as sensitivity analysis) that is dependent to the models, the important variables can be identified via different approaches. Some suggested approaches are briefly discussed in following:

a) For each variable, we calculate the average of its importance score (that is in the scale of 100) coming out of all built models with good quality. The ones with the highest average score can be considered as the important variables. In this approach, the importance score of all variables in every model should be available.
b) We pick the variables that are common in the first 20 important variables of all built models with good quality. Perhaps this approach is less robust comparing to the previous one as there could be a variable ranked second in a model's attribute ranking, but its importance score is less than 50%.
c) If the importance score of variables is not accessible, we calculate the weighted average score (aka the adjusted score) of each variable by multiplying the average score with the square root of its number of appearances in the top important variables of all built models with good quality. For example, let's say we selected five built

models as acceptable, and we only had access to the top 20 important variables of each model. Variable A was ranked in the top important variables of two selected models with an average score of 80, while variable B was listed in the top important variables of three selected models with an average score of 75. Even though variable A has a higher average score, it was not much important in the other three selected models. Therefore, the weighted average score of variable B is 121.24 and would be higher than the adjusted score of variable A which is 113.14.

In regard to building descriptive models, it should be noted that if there is a target variable in the dataset, it should not be included among the input variables. Descriptive models can be built for exploratory purposes to get more insights from the data especially when there is no target variable such as discovering groups of people in the data and defining classes. In addition, they can be built for verification purposes after building predictive models such as investigating whether the identified important variables do play a significant role in differentiating groups of people in the population, and interpreting the – risk or protective – effect of important variables by clustering only using the important variables and checking the majority outcome status in each cluster.

4.3.4 Stage 4: Evaluation

At this point, we report the experimental results obtained from the models. Then, we discuss the quality of models and validity of findings. If the quality of models is satisfactory, we proceed to the evaluation of findings, and if the validity of findings is satisfactory, we publish the results; otherwise, we have to go to the Feedback stage.

The quality of every data mining model is evaluated using various metrics depending on the data mining technique applied to the data. These are some of the common and well-known metrics for each technique:

- **Classification:** correctness accuracy, area under the ROC curve (AUC), sensitivity, specificity, and G-mean (geometric mean).
- **Clustering:** average silhouette width and Dunn index.
- **Association:** lift and confidence.

In what follows, we cover some details on how to evaluate data mining models to obtain reliability.

4.3.4.1 Predictive Models

In evaluating the predictive models, they can be tuned during the modelling process to provide more instances in order to find the optimal model, for example, by changing the maximum depth of trees or number of hidden nodes. Eventually, a confusion matrix is generated separately on the training and test sets to evaluate the predictive model. This matrix demonstrates the frequency of true positive, true negative, false positive, and false negative records in each set.

In case of applying *k*-fold cross-validation during modelling, *k* models are trained for every combination of defined values for the algorithm's tuning parameters. The average accuracy of the *k* models is considered as the estimated accuracy for that particular combination of tuning values. The one with the highest is considered as the best tune and the final model is built using the best tuning values on the entire training set. After we ensure the predictive model has a good quality by examining the models' accuracy (based on the validations done during model building), we should run the model on the test dataset as unseen data. Then, we compare

the quality between the training and test sets using the same metrics. If their quality is very close, then the model is performing well and is stable, even if the accuracy is not high. But if the accuracy on the unseen data is low, it could mean the model is overfit or the patterns for one or more of the classes have not been discovered properly — the model is not reliable. In this case, we need to check the complexity of the model and tune its parameters to avoid overfitting. In the end, the model results on the test set show the actual quality of the built model. In other words, if the estimated accuracy of the trained model is low (usually the final model is the one with the highest validation accuracy among the tuned instances), calculated by averaging the validation accuracy of all rounds in case of cross-validation, that means the model is underfit. If the estimated accuracy of the trained model is higher than the accuracy from test set, that means the model is overfit. The configuration related to the classifier's complexity needs to be increased in the former and decreased in the latter. Figure 4.3 visualizes examples of what underfit, overfit, and just-right models mean.

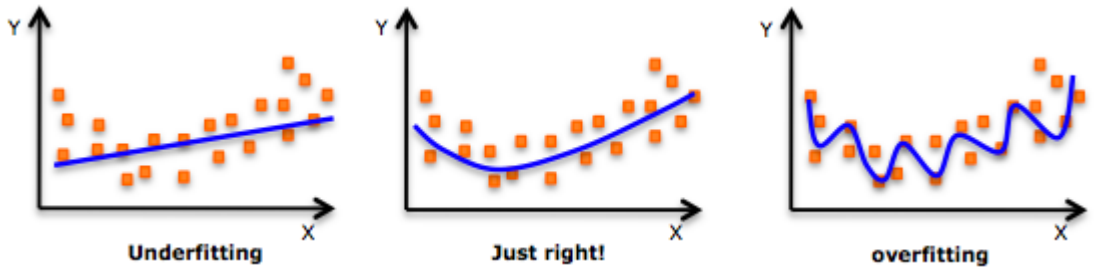

Figure 4.3: Examples of underfit, overfit and just-right models [102]

If the accuracy of the predictive model is low, it could also mean that no appropriate pattern was found, and subsequently, the selected data mining algorithm cannot operate well on the focused data. If the accuracy is low when using various algorithms, either the predictors are not appropriate, in which case a proper pattern cannot be identified that associates with the target variable, or the data have an issue so that appropriate patterns cannot be discovered,

such as the amount of data is low or there are serious conflicts and complexities inside the data.

The quality of classifiers can be measured using various metrics that were mentioned earlier. However, in the health care context, it is important to have models with high sensitivity and specificity in order to find the variables and rules that accurately identify the causes. Furthermore, high sensitivity is essential to identify high-risk patients, and we need to avoid false negatives to the greatest extent. On the other hand, high specificity ensures health care resources are used efficiently and unnecessary health costs are not imposed on the system. In other words, when patients are identified as high risk by mistake, health care resources will be utilized for them to do further investigations when it is not necessary. Therefore, we suggest using metrics that focus on both sensitivity and specificity, such as AUC and G-mean. The accuracy metric can provide an acceptable number while one of the sensitivity or specificity metrics performs poorly, but it is still good practice to measure the accuracy as well. When the accuracy and AUC (or G-mean) have substantial differences, this is a sign that the sensitivity and specificity have significant differences — one very low and other very high. In this case, we should examine the sensitivity and specificity as well to better assess the model's quality.

To compare the classifiers that were built using cross-validation, we should use the test of significance. It is important to use the same seed number when building the models so that the same portion of training set is selected as fold-1 in fitting all models, as it is better to compare the results from the same fold (same data). In other words, it gives us a more precise comparison to compare the performance of models when built on the exact same data records. Basically, the average and variance of accuracy of each fold's model is compared with another

model's average and variance on the same fold of data using t-test, which considers a confidence interval. Comparing the performance of two classifiers has been guided in section 4.6.3 of this data mining textbook [103].

Finally, the attribute-ranking attempts resulting from the predictive models help identify the attributes playing a significant role in identifying the cases. This ranking can be different from one classifier to the other. However, predictors with high rankings in most of the accurate models potentially represent important predictors associated to the analyzed topic. Normally, we cannot expect the classifiers to identify the role of variables in regard to contributing to the risk or being protective of the outcome. In specific cases where simple rules are extracted from the models, it could be possible for the classifiers to judge in this regard. Other possible options could be building descriptive models using clustering and association rules to further elaborate the findings and perhaps be able to assess some contributing variables. Sometimes, it is not easy or even possible to build high-quality descriptive models — especially in high-dimensional datasets — therefore these options might not be available.

4.3.4.2 Descriptive Models

In evaluating descriptive models, we need to consider the characteristics of data in addition to the configuration of the operating machine. For instance, silhouette measurement would require significant memory space to calculate the similarity or dissimilarity between all records if the dataset is large. In these cases, the Dunn index is perhaps a more practical approach.

Other useful information that can be extracted from descriptive models during evaluation is the interpretation of findings. For instance, to find the main characteristics of each cluster, we need to compare the frequency of values in each variable with the frequency in the entire

dataset to be able to identify if particular values are dominant in that cluster. For this, we might need to join the cluster column — which contains the cluster number each record belongs to — with the original dataset that was not preprocessed in order to have access to the original input values (not the normalized ones), which normally have human-friendly meanings, in addition to other input and auxiliary variables that were excluded from modelling. In this regard, if a target variable existed in the dataset but was excluded from modelling, it can also be considered during evaluation to assess the outcome among the main characteristics of the clusters when interpreting them.

4.3.4.3 Reliability vs. Usability

This data mining methodology is designed to provide reliable results when analyzing health administrative data. By "reliability" we mean that the analyst has done nothing wrong both in preparing the data and building models, so the results can be trusted. However, "usability" is another level. When reliable models are built, they either might be usable or not. Models with acceptable performance (in the selected metric) on unseen data that are not overfit are usable. We consider measures above 60% as notable, above 75% as acceptable, and above 90% as high. Below 60% is almost like a random pick and is useless; below 40% is appalling — i.e., a coin flip performs better.

4.3.4.4 Findings Evaluation

It is important to build models with high accuracy in a reasonable time; however, this does not guarantee that the findings are valid and reliable. This reliable methodology enforces significant attention on the contextual preprocessing: the applicable sources of error are scientifically investigated, handled, and addressed using the Contextual Preprocessing Guideline. Therefore, the results, including the model results and extracted findings, are

reliable and can be trusted, even if the accuracy is not acceptable — making the results not usable — as the potential errors have also been reported. However, due to the sensitivity of the topics in the health care domain, it is required that the findings conform with the experiences of clinical experts — i.e., face validity.

The domain expert can help to verify if there are any other explanations related to the findings that do not exist in the literature. For example, the contributing variable could be a confounder. Usually, attribute ranking cannot provide any interpretations per se, in contrast to rules that contain an actionable meaning. Therefore, the domain expert can enhance the interpretations based on their knowledge and experience on other studies in the field.

Note that the input variables were gathered based on the literature review, in addition to the domain expert's suggestions, to make the pool of variables more inclusive. Therefore, we should not verify the findings of the models (i.e., important variables in predictive models and associated variables in descriptive models) based on past works; otherwise, there is no point in building new models to find new rules and hypotheses.

The results can be used and verified in various other ways too:

- Findings of the models, such as rules, either confirm the past works or generate new hypotheses. If our results match the findings in the literature, then this is a good sign that our findings are verified by past works and that they support studies that are compatible with our findings if there are contradictions in other literature. On the other hand, a big advantage of these hypotheses (typically derived from the rules included in the built model) is that they could be multifactorial as we are considering the effect of all possible exposures and their combinations on the outcome, and this is usually

closer to the reality of the cause. These hypotheses can go through individual clinical studies to be examined.

- The predictive models themselves may be used as alerting or decision support systems. Models with fewer input variables and at least acceptable accuracy can perhaps be built using the discovered important variables. The smaller number of required input variables may make the implementation of the system more feasible.
- The outcome can be changed to a related or specialized one (e.g., changing from a group of diseases such as IMIDs to a more specific one such as IBD, or from IBD to Crohn's), and models can be built using the same exposures, then evaluated to see how many of the contributing variables overlap with each other. This could help to understand whether the root causes of the outcomes are in common.

Another verification step could be running a sensitivity analysis on the models. The mathematical models on real-world issues are usually very complex, and the relationship between input and target can not be understood. There could be uncertainty in the input of these models which affects confidence in the output. Therefore, evaluating the impact and contribution of each input variable on the output's uncertainty can be useful to verify the stability of results. For instance, we rank the attributes after building a predictive model. By filtering the important variables one by one and rebuilding the model, we can verify if the quality of model changed significantly or not. If it has, then the filtered variable was really an important one. However, we should keep in mind that all data mining models can be evaluated using multiple metrics that demonstrate the quality or certainty of the models. In addition, the importance of variables in contributing to the model is also measured. Therefore, the findings from the models are reported along with their level of uncertainty as all data mining models

are built on heuristic functions and are fuzzy by nature. Nevertheless, this does not prevent the researcher from considering other assumptions on the input variables based on the identified potential sources of biases in the study — such as errors in classification of values in input variables — and rerunning the model building to examine the changes in the results, if the researcher is keen to further evaluate the reported level of certainty (i.e., the accuracy and attribute ranking) on the models. For example, if one input variable is believed to have a high risk of bias, then most likely it is excluded in the contextual preprocessing step. If another input variable is believed to have low risk of bias and is not excluded, but it is measured as one of the important variables in a predictive model, the variable can be manipulated or even excluded, and then the model building can be rerun and the findings compared. These attempts are considered as the sensitivity analysis on the findings of the model.

External validity is also a concern in reporting and claiming the results of a data mining model. In addition to being internally valid, the data used in that study should represent a population so that the findings can be generalized to an outside population. If the data are based on a specific region in the world, the findings might be highly influenced by race and ethnicity, incidence rates, environmental exposures, and distribution of concerned genes associated to that specific region. Therefore, they cannot be easily generalized to other regions. However, if the health administrative data are population-based and collected independent of the study, this increases the external validity [104]. It is highly suggested to set the scope of results and findings based on the selected data to increase the confidence and reliability of published results. Note that limiting the scope (i.e., the source population) extensively to match the study population would reduce the risk of selection bias, but at the same time, it will reduce the external validity of results too. In other words, the source population will become very specific

and might not be able to generalize the results to the original population we intended to study. For instance, if we set the scope of a study to diabetic patients visiting a hospital only in the month of February, and do not select samples from visits in other months of the year, then the findings might not be generalizable to all diabetic patients living in the hospital's region.

Lastly, other verification attempts such as retesting the results in a sub-cohort or statistical hypothesis testing can also be conducted for further confidence in the discovered information from the data mining models.

4.3.5 Stage 5: Feedback

At the Evaluation stage, if the quality of models is satisfactory, we proceed to the findings evaluation; otherwise, we have to take the feedback and return to the Modelling stage again to improve the quality of models. If the findings are satisfactory, we publish results and can go to incremental learning mode, if needed; otherwise, we must return to the Preprocessing or Modelling stages depending on the feedback.

The models usually need to be improved for various reasons, with the goal of producing more usable and up-to-date results. These improvements can be done after learning from every experiment by changing the parameter values, replacing algorithms, filtering attributes, or even updating the dataset.

As demonstrated in Figure 4.2, there will be feedback from the evaluation to continuously improve or update the results. This feedback will ensure the models are sustainable and updated. If the measured accuracy of models built using the selected attributes is acceptable, the models can be incrementally rebuilt. However, if the accuracy is below the acceptable threshold (which varies depending on the technique and the metric), or the findings

demonstrate that an attribute does not play a significant role in detecting the hidden patterns, or the domain expert suggests including an attribute in the analysis that is related to one of the significant attributes resulting from the previous models, the feedback takes the process back to the preprocessing stage. Note that it is assumed that all relevant and available attributes were gathered in the first stage, therefore it is not necessary to go back to that stage in order to apply the attribute-modification feedback.

Another applicable idea is that when stable models are achieved and certain conditions are satisfied, the model is upgraded to a version that can learn from new batches of data added to the dataset in order to keep producing updated results. Following the well-known concept of "lifelong learning" introduced by Thrun [105], the capability of learning, retraining, and using knowledge on an ongoing basis to improve the models' performance can be implemented. The goal is to retain learned knowledge and transfer that knowledge when updating models.

An instance where feedback is given to the preprocessing stage is when there are many variables, and it would be very difficult to achieve a high-quality clustering. Therefore, one suggestion is to pick the top five important variables, or apply PCA or linear discriminant analysis (LDA) to reduce the dimensions. If the latter is chosen, the clustering is done on a few combined variables. However, if the quality of the model is good, the cluster number each record belongs to is determined. Then, we assess the main characteristics of each cluster using the original variables — not the new combined variables. In other words, the combined variables generated using PCA or LDA are an intermediary step in finding high-quality distinct clusters within the dataset. We can also use the new variables from PCA or LDA to plot the clusters in two or a few dimensions.

Lastly, it is again expected that the model is built in a reasonable amount of time. If not, demanding feedback to the modelling stage would be to boost the models with scalable solutions such as big data analytics.

5 PREPROCESSING GUIDELINES

In this chapter, we present the preprocessing guidelines developed as one of the key contributions of this thesis. These guidelines are applicable to the second stage of the reliable data mining methodology on health administrative data that we designed in this research study. In other words, these are the guidelines that attempt to take the analyst step-by-step through the preprocessing stage when using health administrative data to build data mining models.

5.1 Technical Preprocessing Guideline (TPG)

As we intend to build data mining models, we suggest applying the following steps to the selected health administrative datasets to make them suitable to run data mining techniques and increase the reliability of results. This guideline will direct the investigator step-by-step to prepare the data for both predictive and descriptive modelling by designing a preprocessing plan and implementing the actions accordingly.

5.1.1 Step 1. Dataset Creation

Applies to

- Both predictive and descriptive modelling

Cause

Health administrative data are collected in various structures and formats that might not be suitable and understandable by the data mining techniques. For example, each record in a gathered health administrative dataset might represent a visit, and there could be multiple

diagnosis columns per visit, most of which have null values. In addition, perhaps not all requested features and patients are available in one available dataset.

Note that it is possible that one main dataset is created, and then multiple versions are derived from it to experiment with various analysis goals. If this is the case, then the next steps of the technical preprocessing should be planned out and applied by considering to which final dataset each variable belongs. For example, the main dataset may contain over a hundred variables, thus we think it requires dimensionality reduction. However, the plan may be to prepare ten derived datasets that each contain less than twenty variables and are not high-dimensional.

Actions

Run the following action items:

1. **Base dataset:** if possible, select the dataset that contains the focused population and requested variables more than other datasets as the one to start with.
2. **Dataset structure:** a two-dimensional tabular dataset needs to be formed in which each record represents one patient. All features related to the patients should be added as separate columns in this dataset. Therefore, the columns of this tabular dataset represent the input and target variables.
3. **Inclusion/exclusion criteria:** apply the inclusion and exclusion criteria defined in the data creation plan to include all the existing valid patients.

5.1.2 Step 2. Feature Construction

Applies to

- Both predictive and descriptive modelling

Cause

Data mining techniques attempt to identify patterns in data and discover the association between a wide range of input variables and possibly a target. But, health administrative data usually include only necessary and particular pieces of information captured as a matter of running a health administration system. As the data are not collected for predefined research questions and are being used for secondary purposes, not all required information is captured directly or in a suitable structure. However, some information can be created using the available health administrative data to represent the concerned risk factors.

For example, to indicate whether a patient was truly diagnosed with a disease in a dataset, a validated diagnosis algorithm associated with the incidence of the disease is used. These algorithms usually check the number of specific diagnosis codes recorded for the patient via different health providers and in various age categories. As each patient visit is normally captured in a separate record, the records belonging to each patient need to be explored to construct the required information in a structure that is understandable to the techniques.

Actions

Generate all required variables that represent the focused risk factors by linking the available data sources, combining related records, calculating new raw information, and categorizing the values into groups.

5.1.3 Step 3. Exploration and Planning

Applies to

- Both predictive and descriptive modelling

Cause

Now that we have a two-dimensional tabular dataset, the included variables and records need to be investigated and then a plan developed on how to properly preprocess them. The preprocessing plan will be filled out as we proceed with the next steps of this guideline. This plan helps us significantly in designing the preprocessing actions and implementing them in the selected platform.

We introduce our best practice in the following; however, the investigator can design the preprocessing plan in any format that best accommodates their conditions and procedures.

One of the outcomes of this step is to determine each variable's role which could be either input, target, auxiliary, intermediary, or none. The input and target variables will be listed as per the modelling purposes. The auxiliary variables that are intended for modelling will still be preserved in the dataset for further investigations on the results, but no further preprocessing will be applied to them. The intermediary variables that were used to link source datasets or construct new variables can be eliminated from the dataset. Hence, from this point, we only need to focus on and preprocess the input and target variables. Finally, we must keep in mind that the initial dataset can be preprocessed to provide multiple final datasets to run experiments with different goals, and each variable could have a different role in any of the intended final datasets. Therefore, we need to specify to which final dataset and role each variable should belong. It is likely that a given variable belongs to one final dataset and not in the others. So, its role in the other final datasets will be none.

Note that if the dataset is being prepared for descriptive modelling, there still could be target variables. If we have a specific target variable — intended for a supervised learning — in the dataset, it will still be marked as "target" in the plan and will act as an auxiliary variable. It should not be part of the input variables. In this case, the target variable can be used to further

investigate the unsupervised/undirected results. For example, a clustering model can be investigated with respect to a target variable to verify its status in each cluster. This is called a directed clustering.

Actions

A good practice to better apply this guideline is to fetch the frequency of values in all available variables and summarize them in a spreadsheet at the point where the dataset is created and all concerned variables are constructed. Then, the basic information and statistics of each variable are captured in this spreadsheet including the description of the variable, purpose or associated risk factor, data type (continuous, categorical, date), data subtype (binary, nominal, ordinal, real, natural), variable role (input, target, auxiliary, intermediary), frequency and percentage of missing values, number of distinct valid values, percentage of most frequent and least frequent values, whether it contains inconsistent values, and whether the missing values have a valid meaning. The required actions can be planned out on this spreadsheet, such as removing as duplicate, removing as intermediary, removing due to high missingness, removing as correlated, removing as unary or almost unary (after excluding the missing values), requires imputation, and requires binning. This includes filtering variables as a matter of preparing multiple versions of the dataset to build models for different goals.

Note that variables that have been included for further investigation purposes and are not intended for the modelling stage can still be listed in the spreadsheet, but do not need any statistics investigation and preprocessing actions. Furthermore, at this point:

- the intermediary variables can be eliminated from the dataset.
- the data type and subtype of variables can be fixed based on the selected platform.

- the input and target variables can be listed for modelling purposes, and also the next preprocessing actions are applied only to them.

5.1.4 Step 4. Data Partitioning

Applies to

- Only predictive modelling

Cause

If we are planning to build a predictive model, we need to partition the data into training and test sets as soon as the initial structure of the dataset is prepared. Even though the dataset has not been properly preprocessed yet, this data split is essential to build a reliable predictive model. As a result, the upcoming preprocessing steps will be planned based on the data that exists in the training set and then applied to the training set. Later, the same preprocessing models or parameters that were developed based on the training set will be applied to the test set. This is important, especially if the next preprocessing actions depend on the data (e.g., standardization and PCA). For example, if you are centering your data (subtracting the mean) then you should only calculate the mean of values using the variable in your training data, and then subtract this mean from the values of the variable in both your training and test data. You should not calculate the mean of the variable on your test data. This is because we are considering the test set as our unseen data to evaluate the predictive model and we need to keep the experience as real as possible. In other words, the training set is present data and all we know about, while the test set is future data and unseen. The details of the test set should not be exposed to the model, otherwise it will not provide a fair test of the model. If the entire dataset is preprocessed together, that means the portion that will be split later as the training set will include information — such as the range, outliers, and distribution — from the portion

that should be considered as unseen data. Hence, there will be an indirect information leak and the predictive model will be trained in a way that is already aware of the unseen data's statistics. This could cause overestimation of the model's accuracy, causing the model to not operate with the same quality on actual unseen data. The actual unseen data are preprocessed based on the parameters developed from preprocessing our experimental dataset. Note that the same issue is applicable to the validation set, which will be discussed in the "Resampling Data" step. Ideally, the validation set should be preprocessed using the preprocessing parameters from the portion of training set that fit the model; however, this is not always possible or feasible to account for.

In addition, we need a properly preprocessed portion of the dataset retained as the test set to be able to correctly evaluate the quality of the trained model (aka the holdout method). Even though the training set will be resampled to generate a validation set and estimate the accuracy accordingly, the accuracy retrieved from running the model on the test set — the unseen data — suggests whether the model is performing well with regard to not being overfit or underfit. An overfit or overtrained model is typically one that is too complex in structure and has learned all the details in the training set, so it is no longer flexible enough to do a good job on new data. On the other hand, an underfit or undertrained model is too simplified and cannot recognize the various patterns in the data. For example, linear regression is at high risk of underfitting, while a high degree in polynomial regression, a high number of nodes and levels in decision trees, and a high number of hidden layers and nodes in neural networks could cause overfitting. The complexity of the model can have a relationship with the size of the data, but the data volume per se does not cause the model to be overtrained. As the volume of data increases, the room for the model to become more complex and identify more rules while still

not overtraining increases too, and vice versa. If the predictive model is underfit or overfit for any reason, this can be identified by comparing the accuracy of the model between the training and test sets. This issue will be discussed further in the evaluation stage.

Another challenge concerns untrained values. Even though this issue might seem to be related to data partitioning, it can happen even without partitioning. If the test set or new data contain values in categorical variables that are not included in the training set, the model simply cannot consider them as it has not been trained for those specific values and does not know about their relationship. Nonetheless, these categorical values will most likely be ignored during preprocessing as a matter of one-hot encoding their variables based on the training set's information; this is further discussed in the "Categorical Variables" step. Data partitioning could also cause this issue to happen if it was not done very well (i.e., proper randomization) or one rare value in a categorical variable is only included in the test set. Therefore, the training set will not contain this value, and the model will not be trained for it. With regard to continuous variables, there should not be any problem. Even if the new value is greater than the maximum value existing in the training set or lower than the minimum value, this value will be preprocessed to a new value outside of the boundary (e.g., 0 and 1 if the data were normalized) in the test data. The equations could still understand the number and calculate the distances properly.

Actions

It is suggested to randomly split the data into training and test sets with a 70/30 percentage proportion. If the dataset is not large, the split can be shifted toward 80/20 percent so that there is more data included in the training set. Random selection of records should provide the same

ratio of classes in both training and test sets as the ratio in the original dataset; otherwise, the selection has to be done in a way to preserve this ratio in both training and test sets.

5.1.5 Step 5. Data Cleaning

Applies to

- Both predictive and descriptive modelling
- If the dataset is partitioned, the actions are planned only based on the training set, then applied to both training and test sets

Cause

Data mining techniques expect the data to be well-prepared and consistent. Health administrative datasets are collected as a matter of running health administration systems and from multiple sources. Real-world data typically contains noise, inconsistency, and various formats. Therefore, it needs to be cleaned before modelling. Note that valid outliers or anomalies should remain in the data.

The inconsistency can exist in missing data as well. Missing data in some variables can be in the form of null values, while in others it can be displayed with empty string, whitespace, negative one, etc. This issue might not even exist in the first place in the raw data, but instead shows up when the data are transferred to another platform. For instance, when a character variable containing missing values in SAS is imported into R, it will be read as an empty factor level and not an NA (actual missing value). However, it is expected that missing values are all transformed into a consistent null value, depending on the platform, so that they can be handled properly in the “Missing Values” step by imputation or other methods.

Actions

Examine the frequencies retrieved in the "Exploration and Planning" step to check the following items and fix the values in each variable accordingly:

- **Case sensitivity:** text values that represent the same concept or meaning should have the same case, for example, all in lowercase; otherwise, the models might consider them as separate values
- **Blank cells:** cells that contain an empty string, white space, or null values should be harmonized to one value representing blank, that is, missing value
- **Format:** every variable should use the most appropriate data type and follow one specific format; related issues include but are not limited to:
    - numbers stored as text
    - different date and time formats
    - multiple values for one concept, for example, using "M" and "1" to show male patients in the sex variable
    - different measurement units, for example, temperatures captured in Celsius and Fahrenheit in one variable
- **Outliers (anomalies) and extreme values:** outliers and extreme values should not be removed unless they are invalid values (noises) or skews the models significantly that cannot find patterns among most data. For example, 99% of records are grouped in one cluster because of the outliers, while there are serious differences in characteristics of potential groups within that one cluster.
- **Noise:** any other inconsistencies and unexpected values are considered as noise, which includes but not is limited to:
    - misspelled text values need to be fixed

- out-of-context values when comparing to other distinct values, for example, having “M”, “F”, and “125” in the sex variable, need to be omitted and considered as missing values
- out-of-range outliers, by checking the minimum and maximum values, for example, an age of -100, need to be omitted and considered as missing values
- leading or trailing white spaces need to be trimmed

5.1.6 Step 6. Missing Values

Applies to

- Both predictive and descriptive modelling
- If the dataset is partitioned, the actions are planned only based on the training set, then applied on both training and test sets

Cause

One of the most difficult issues during preprocessing is eliminating missing values as there is usually no good solution for it. Most data mining techniques cannot handle missing values. Health administrative data are gathered from multiple sources that usually contain incomplete information shown as missing values.

Missing data can be in various formats including null values, empty strings, or whitespaces. In the “Data Cleaning” step, this issue should have been handled already, otherwise it could cause trouble and inconsistent behavior in the next steps.

Actions

Input Variables

Use the following strategies in the listed order to handle the missing values in input variables, depending on the condition:

1. **Keep:** retain them if you intend to use decision trees or naïve Bayesian models that can handle missing values as a separate category.
2. **Replace:** use a specific numeric value if the missing values are valid and have meaning, for example, death flag and no income.
3. **Remove variable:** omit the variable if the amount of its missing values is too high (e.g., >60%).
4. **Truncate record:** if the target variable of a record is missing, it should be removed; in addition, if a record contains many missing values (e.g., more than one fourth of cells among only the input variables altogether — after removing the variables with high missingness) and is not handled using any of the above solutions, it can be truncated. Note that this strategy is suggested in cases where enough data (e.g., at least tens or hundreds of thousands of records) remains for building data mining models. If this option cannot be applied, the advanced imputation methods might also not respond well as they cannot match the records with high missingness with similar ones.
5. **Imputation:** if the number of missing values is low, imputation methods can be used to replace the missing values with values that would most likely make sense based on similar records and also preserve the distribution and other statistics as much as possible. Depending on the software environment and packages being used to do the implementation, imputation might need to defer to after "Step 9. Data Normalization" so that one-hot encoding of categorical variables and normalization of continuous variables are done. The imputed datasets should be verified so that the new data are

consistent in terms of the variability of values by checking the mean, variance, and distribution. The verification checks are listed below. There are two categories of imputation methods:

a. ***Basic imputation methods***, such as mean, median, and mode. These methods are not recommended as they will destroy the variability of values and distort the distribution. However, if the advanced methods are not possible for any good reason (lack of resources or implementing a test run), the simpler methods can be a quick alternative resolution.

b. ***Advanced imputation methods***, such as k-nearest neighbors (KNN), regression (or single) imputation, stochastic imputation, and multiple imputation. In cases where the dataset is high-dimensional, using sophisticated methods (e.g., KNN and multiple imputation) can be significantly time-consuming, but they usually provide a better estimation of variances. If there are not enough resources available to speed up the process, simpler and faster methods (e.g., regression or single imputation) can be better choices and still produce an acceptable quality of synthetic data. Further discussion about some of the imputation methods can be found in the following section entitled “More Information on Imputation”.

c. ***Imputation verification:*** The dataset should be compared before and after imputation (using PROC COMPARE, MEANS, or FREQ in SAS) to check the following items:

    - The value differences are only on the missing cells (i.e., the number of missing differences between two datasets in each variable should be the

same as the number of unequal values between two datasets in the same variable).

- No missing data exist after imputation.
- The frequency of values between the dataset before imputation (excluding the missing data in percentage calculations) and the dataset after imputation is almost preserved. The difference of some variables can be shown using histograms.
- Mean and standard deviation of continuous variables before and after imputation is almost preserved. The difference of some variables can be shown using boxplots.

Please pay attention that the order of applying the above strategies is important and could significantly help in preserving the quality of data, and consequently the useful information in the data, so that we do not lose records as much as possible.

Target Variable

Depending on how the target variable is formed, the following options are suggested to address missing values in the target variable:

- If the target variable has been constructed during the study's preprocessing stage, the criteria need to be revised so that the target variable does not end up with missing values.
- If it already existed in the study databases: a) the values should be replaced in case the missing values have a valid meaning, b) the records should be truncated if they cannot be fixed as there should not be any missing values in the target variable and they also cannot be imputed.

Other Variables

Do not worry about missing values in non-input, non-target variables that exist in the dataset only for further investigations on the results; there is no need to apply any actions on these variables.

More Information on Imputation

It is a common misunderstanding that imputation methods should replace the missing data with "real" values. However, the main purpose of applying an imputation method is to provide imputed values in a way that can correctly reproduce the variance and preserve the distribution of values if our data did not have any missing data.

5.1.7 Step 7. Categorical Variables

Applies to

- Both predictive and descriptive modelling
- If the dataset is partitioned, the actions are planned only based on the training set, then applied on both training and test sets

Cause

Some variables, usually categorical ones, are connected to each other and provide information related to the same concept based on how the data collection procedure was designed in the various source health administrative databases. Perhaps these variables can be merged if the study is not interested in the distinctions between them. The same information may be captured via different means, methods, and even from separate sources, such as breastfeeding during the hospital stay, at discharge, and at home. Feeding too many variables to data mining techniques will decrease the stability of data and accuracy of models due to sparse data points.

On the other hand, categorical variables that have too many distinct values will increase the complexity of calculations in the model and make it difficult to discover appropriate patterns. In addition, some continuous variables have standard cut-offs, such as low, medium, and high blood pressures, or top-coding ages over 65. When the study is more concerned about the relationship of the known groups or ranges of values, the values can be binned to reduce the number of distinct values. Otherwise, it might not be a good idea to bin the values thoughtlessly and without any good reasoning as binning into too few categories could cause loss of information, while binning into too many categories could cause confusion in finding associations.

Finally, all data mining algorithms are based on mathematical equations; therefore, all values should be in numeric format to be able to run the calculations. The continuous variables are fine as they already have numeric values. However, categorical variables will face problems in this regard. One solution is to replace the alphanumeric values with numeric ones. Even if they are already in numeric form — for example, 1, 2, 3, and 4 — each represents a distinct group. The equations will incorrectly assume that group 4 has a distance of 3 units with group 1 and group 3 is closer to group 4, while considering a distance between groups might not even make sense — what is the distance between red, green, and blue? So, the other solution is to break down each categorical variable into multiple binary dummy variables (with values 0 or 1); each dummy variable shows the presence of the corresponding value. Only one of the dummy variables will be true, while the other related dummy variables are false. Therefore, this process is called “one-hot encoding.” In cases where a categorical value is a combination of some other values, it does not need a separate dummy variable as the independent values are sufficient. For example, a categorical variable representing the feeding status of a baby

can have three values: breastfeeding, formula feeding, and mixed. When this variable is one-hot encoded, the dummy variable for mixed feeding can be eliminated and instead both dummy variables receive a binary true value. Lastly, there is a concern about ordinal variables as there is a natural or logical sequence in the distinct values. The first solution will preserve this sequence, but this may impose an unrealistic distance between them (e.g., the distance between BSc, MSc, and PhD degrees can be different from various perspectives, such as the number of years, amount of effort, academic value, etc., even if it is possible to quantify them). On the other hand, the second solution will lose the sequence, while not misleading the equations by injecting unrealistic distances between the values. Note that if the categorical variable only has two distinct values, one-hot encoding is not needed and simply transforming the values to 0 and 1 would be sufficient. Also, as the test set will be one-hot encoded based on the available distinct values in the training set, it is possible — in real-life scenarios as well — to have a new distinct value in the unseen set which will not receive any dummy variable as the model is not even trained to handle this new category. The value of all related dummy variables will be false.

Actions

The definition and values of all input variables should be examined and the following strategies applied, when applicable, in the listed order:

- **Numeration:** we need to have all variables in a numeric format. The continuous and binary variables should be preserved or transformed, if needed, into numeric format. The ordinal and nominal variables should stay in a categorical format (e.g., factor levels in R) and will be taken care of in the following actions. The only exception is

the target variable which should stay as a categorical variable unless a regression model is going to be built.

- **Aggregation:** merging the variables, usually categorical, that are conceptually related to each other to additional variables with overlapping pieces of information. A measurement unit or new set of values that best fits the concept of the variable, preferably using fewer distinct values, should be used.
- **Categorization (or Binning):** putting numerical or even categorical values into a set of defined groups, especially if there are many distinct values (e.g., >100) and there are standard cutoffs or categories already defined for the variable. The criteria should be defined based on a rationale so that each bin contains conceptually related values. Otherwise, this action could cause confusion in the model and the underlying pattern cannot be discovered. In addition, if the target variable has lots of distinct values and classifiers are intended to use, the values should be binned.
- **One-Hot Encoding (OHE):** all categorical variables who have two distinct values should transform into 0 and 1 values, while the ones with more than two should be broken down into multiple binary dummy variables wherein each dummy variable represents the presence or absence of one distinct value. If the categorical values demonstrate the possible variety of categories (e.g., age groups of receiving the first flu shot) and there is no categorical value that demonstrate the presence of any type (e.g., never had a flu shot), it is suggested to create a separate dummy variable to include the impact of belonging to any of these categories — despite their differences — in contrast to not belonging to any of them. Normally, missing values might act as this type of categorical values in such variables. If they are replaced with a valid value

in “Step 6. Missing Values”, this concern will get handled automatically when OHE is done.

5.1.8 Step 8. Feature Reduction

Applies to

- Both predictive and descriptive modelling
- If the dataset is partitioned, the actions are planned only based on the training set, then applied on both training and test sets

Cause

Redundant variables do not provide much information; in contrast, they will confuse data mining models in finding the patterns and the optimum set of important variables. Health administrative data are collected from various sources; therefore, the data could have irrelevant and duplicated variables.

Furthermore, high-dimensional datasets (i.e., tens and hundreds of variables) could face a curse of dimensionality in which the data are sparse and the records become too dissimilar. Models will not be accurate enough to find patterns within the data.

Actions

Remove the following input variables to just return a subset of features:

- **Derived variables:** the variables that are calculated using other existing independent variables in the data.
- **Unary (zero-variance) variables:** the variables with one distinct value, excluding the missing values.

- **Almost-unary (near-zero-variance) variables:** the variables with more than 95% unary value, excluding the missing values. In addition to the very low information available in these variables, they can become unary when resampling happens if the data size is small and eventually cause failures in predictive modelling algorithms.
    - *Exception:* if the dataset is large (in which an almost-unary variable most likely will not become unary due to resampling during training) and the target variable is highly imbalanced (which could be suspicious of having a relationship with the almost-unary variables), almost-unary variables can remain.
- **Unique variables:** the variables containing unique or direct identifying values such as name, phone number, and ID number.
- **Highly correlated variables:** the variables that are highly correlated (e.g., the absolute correlation value >0.9) should preferably be reduced to a single variable, unless the expected model — for example, partial least squares (PLS) — benefits from including correlated variables. Another option is to run a dimension reduction technique such as principal component analysis (PCA), linear discriminant analysis (LDA), canonical correlation analysis (CCA), or non-negative matrix factorization (NMF) to reduce the number of input variables. Otherwise:
    a. the input variables that are duplicated or highly correlated to each other (and their correlation is conceptually valid) should be reduced to a single variable
    b. the input variables that are synonym or highly correlated to the target (and their correlation is conceptually valid) should be removed

Note that if there are many missing values in the data and the imputation has been deferred, the correlation calculation should also be postponed to after imputation so that it does not have to ignore rows with missing values, otherwise it could provide unreliable correlation values between columns as it calculates using a small number of records.

5.1.9 Step 9. Data Normalization

Applies to

- Both predictive and descriptive modelling
- If the dataset is partitioned, the actions are planned only based on the training set, then applied on both training and test sets

Cause

Every variable has a different range of values. Even though this is totally understandable due to the purpose and meaning of those numbers, most data mining algorithms will compare the values across variables based on Euclidean distance. As there are no common grounds, the variables with large numbers (such as salary) will be dominant. Therefore, it is important to normalize the values. This action is also called “Feature Scaling.”

In algorithms that are not based on Euclidean distance, normalizing the data will still be beneficial as the model will converge much faster. As an example, decision trees are not based on Euclidean distance, but they will run for a shorter time when the data are normalized (i.e., the features are centered and scaled).

Actions

Apply a normalization method, such as one of the following, to all variables except the target variable to prepare common grounds for comparison by scaling them:

- **Rescaling (Min-Max Normalization):** normalizing the data in each variable by subtracting all values with the minimum value in that variable and dividing them by the length (the difference between the maximum and minimum values in that variable) to bring the values in the range of 0 to 1.
- **Standardization (Z-Score Normalization):** standardizing the data in each variable by subtracting all values with the average value in that variable and dividing them by the standard deviation of values in that variable to provide an average of 0 and standard deviation of 1.

5.1.10 Step 10. Data Balancing

Applies to

- Only predictive modelling
- Only the training set

Cause

In order to build predictive models, classifiers should be fed with balanced datasets. A balanced dataset contains equal portions of records associated with each class (aka distinct values in the target variable). If the dataset is not balanced, the predictive model will be biased, that is, it can predict the majority class much better than the minority class.

Actions

Use one of the following methods to make a balanced version of the prepared datasets:

- **Undersampling:** randomly pick records from the majority class to have the same number of records as the minority class. This method is simple and effective but lowers the accuracy as the amount of data is reduced. It is suitable when the minority class size is not very small compared to the majority class size and the data size is large.
- **Oversampling:** randomly select and duplicate records from the minority class to have the same number of records with the majority class. This method can overfit the classifiers. It is suitable when the minority class size is not very small compared to the majority class size and the data size is not large.
- **Matched case-control:** for every case, pick a record from controls that match in particular attributes, usually demographics such as sex and date of birth. This method provides better comparing record pairs so that the main differences between cases and controls are distinguishable; however, the matching attributes are arguable and usually selected based on the literature. Also, this method is suitable when the minority class size is not very small compared to the majority class size, otherwise it could lower accuracy if the cases are the minority or overfit the model if they are the majority.
- **SMOTE (Synthetic Minority Oversampling TEchnique):** create records that are interpolations of the minority class and also undersample the majority class. This is a better oversampling method than the basic one. If using cross-validation, make sure to balance the training data in each iteration instead of balancing the whole dataset before cross-validation to avoid overfitting.

5.1.11 Step 11. Resampling Data

Applies to

- Only predictive modelling

- Only the training set

Cause

To build predictive models, a segment of the training set is considered as the validation set (aka temporary unseen data) to estimate the model's performance. There are various resampling methods that can be used in partitioning the data to build predictive models, such as data split (e.g., a 70/30 split), bootstrap (taking *b* random samples from the dataset with replacement), *k*-fold cross-validation (CV), and repeated *k*-fold cross-validation.

It is assumed that the repeated *k*-fold CV provides the best estimate of the model's performance, while data split is the worst as the chance of error in estimating the accuracy is higher. However, their computational cost is exactly in reverse. As we are dealing with health administrative data, the volume of data is usually large, and at the same time, the accuracy of findings is sensitive as it is being used for health-related purposes. So, we cannot use a method that estimates the accuracy poorly, while at the same time, it should be feasible.

As mentioned earlier, the whole purpose of this data resampling is to estimate the accuracy of the trained model as best as possible. The final model is trained using the entire training set using the tuning values with the highest accuracy estimation. However, we should keep in mind that the validation set is a portion of the training set, which has been preprocessed as a whole. Usually, the validation set cannot be preprocessed based on the preprocessing parameters from the portion of the training set that fit the model — even though ideally that is preferred. Therefore, the validation set is not perfect unseen data and the variances in its data have been leaked to the model. In addition, there are usually various attempts at building and validating the model. Therefore, the accuracy retrieved from the validation set is only considered an estimate that is slightly biased and might be optimistic in most cases. If the

dataset is small, especially with categorical variables that contain a lot of levels, the estimate is much more biased. Overall, resampling the data right after partitioning them into training and test sets is possible, so that only the actual training set (which excludes the validation set) is preprocessed and then the validation set is preprocessed using the parameters from the training set — exactly what we do with the test set — to provide more accurate estimates of the model's performance. However, if a resampling method like cross-validation is selected, along with the model-tuning process happening during modelling, this means the preprocessing steps should be repeated many times. If there are costly steps in the preprocessing actions, such as advanced imputation, the overall computational time would be enormous and not feasible. Furthermore, we still have the test data that can be used to evaluate the model properly, so we do not have to worry too much about the precision of the estimates. Therefore, our general suggestion is to do resampling at the end of preprocessing as health administrative data are usually large.

Actions

It is recommended to use the *k*-fold cross-validation resampling method when building a predictive model on health administrative data to achieve a better result on the performance estimates and computational time. More specifically, it is advised to use 10-fold cross-validation as it has been observed that the precision of estimates could drop quite significantly when downgrading to 5-fold cross-validation.

5.1.12 Step 12. Runtime Filtering

Applies to

- Both predictive and descriptive modelling

- If the dataset is partitioned, the actions are planned only based on the training set, then applied on both training and test sets

Cause

As preprocessed health administrative datasets normally still end up with lots of variables, it is a common practice to filter out some of the variables before building a new model based on the results of the previous runs for various purposes, such as experimenting with the impact of different variables or excluding the least important variables from the previous runs. In this case, it is not necessary to build multiple versions of the dataset by removing the concerned variables each time. Creating a new version of a health administrative dataset is usually very time-consuming and storage intensive.

Actions

At runtime, when the dataset is being fetched from the disk or database, exclude the input variables whose impacts on the model are not going to be experimented with in this run.

5.2 Contextual Preprocessing Guideline (CPG)

Health administrative data are primarily collected routinely as a matter of running health systems including clinical care, billing, and administration. Even though administrative databases are not created to address research purposes, they can become available for retrospective research studies as the secondary use of data. Therefore, research with health administrative data typically risks misclassification bias, confounding, and errors due to missing data that need to be properly considered during study design [104].

Most sources of potential bias can be minimized if they are considered when the study is being designed, especially in the data and preprocessing stages of the methodology when the data

are being selected and prepared. Nevertheless, some biases could still stay in the data and affect the results interpretations. Therefore, potential biases should be identified and discussed along with the results.

The following steps examine the possible issues related to the scope of this methodology, which is applying data mining techniques to health administrative data.

5.2.1 Step 1. Selected Source Databases

Risk

- Selection bias: Ascertainment bias
- Information bias: Misclassification bias

Cause

The coverage and quality of selected source databases could impact the validity and generalizability of the study results. The health administrative data used in a research study (i.e., the study population) are retrieved from one or multiple administrative databases. Administrative databases are usually population-based with broad geographic coverage, large sample sizes and long follow-up periods [104]. However, the selected health administrative databases (i.e., the database population) might not include all individuals the study is focused on (i.e., the source population). This could introduce selection bias to the study. The more the database population covers the source population, the less selection bias is included in the study. Figure 3 demonstrates the hierarchy of these types of population.

Even though population-based databases have high coverage, they normally do not include all the individuals from the source population. In addition, the selected databases could be

population-based, but the coverage assumed as the source population (e.g., an entire country) is much broader than the defined coverage for the databases (e.g., one province).

The data collection method in each source database demonstrates which population is focused on and what information is captured in that database. Even though the data recording in administrative databases is independent, which minimizes non-responder, recall, and selection biases [104], the procedures taken to collect the information could affect the validity of the variables. Note that the database could have low risk of selection bias in its own context, but when compared to the source population required for the research study, it could contain selection bias for that specific study. For example, the records of all hospitalized diabetic patients in a city are gathered in a database, and for that purpose, it is a complete database. Our study is using this database to analyze all diabetic patients in that city. Of course, not all diabetic patients in that city have necessarily been hospitalized, so there is a mismatch between the source and database populations which leads to selection bias in our study.

As the study population is retrieved from the selected source databases, it is important to know which individuals are included in these databases, the key variables, what each record in their datasets represents [106], and how the data were collected and measured.

Actions

Identify Risk

Prepare a description for each selected source database which specifies the main reason for creation and how the data were collected, and then:

- compare the coverage of database population with the source population to identify the amount of difference/mismatch and assess the existence of selection bias; and

- evaluate the procedure of data collection by examining the existence and significance of quality control, validation, and auditing during the procedure in order to identify the chance of misclassification in these databases.

Mitigate Risk

If the coverage of the database population compared to the source population is very limited and the risk of selection bias is assessed as high, suggestions include:

- go back to the previous stage of the methodology and select databases with higher coverage on the source population or collect supplementary data; or
- if the missing portion of the source population is known (e.g., there are strong evidence that more records from a particular segment of population with known demographic distributions should be added), build synthetic data and impute the unknown details as part of addressing missing values in the technical preprocessing.

If there are serious concerns on the validity of data as a matter of flaws in the data collection procedures, or significant categories of key variables are missing, it is recommended to go back to the previous stage of the methodology and select databases with higher reliability in data collection (such as robust quality controls, validation, and auditing) and variable collection.

The level of risk on each source of bias discussed above should be reported along with the study results, especially if the mentioned suggestions are not applicable and the risks are high.

5.2.2 Step 2. Criteria of Study Population

Risk

- Selection bias: Inclusion and exclusion biases, matching bias

Cause

When cases and controls are selected with different mechanisms, selection bias could happen. Especially if they are picked from different sources — for example, cases being selected from ambulatory or hospital-based sources and controls selected based on feasibility — the systematic differences would increase the risk of bias [104]. These items should be considered during the stage of defining inclusion and exclusion criteria to select the study population from the database population.

In addition, some individuals may need to be excluded from the study population if they do not have enough follow-up period (due to lack of access to them or death), they have corrupt data, or other reasons. This could also be a result of errors in data linkage in which multiple datasets are joined into one and not all samples (or associated records) exist in all datasets, resulting in some samples being dropped.

Finally, if the controls are selected from the same database population to match the cases based on specific demographic characteristics, again there is a risk of selection bias depending on the defined matching criteria.

Actions

Identify Risk

Review the inclusion and exclusion criteria which formed the study population and created the study dataset at the previous stage of the methodology to verify the following items:

- The included individuals (cases and controls) in the study population should properly represent the source population. If relevant individuals available in the database population are not included, there are chances of inclusion bias in the study.
- The invalid individuals are removed based on the exclusion criteria due to various issues such as a limited follow-up period or corrupt data. However, if certain characteristics are lost due to the exclusion criteria, this can introduce exclusion bias to the study.
- If the controls are selected based on matching criteria, they need to be carefully verified as there is a chance of matching bias.

Mitigate Risk

If inclusion or exclusion bias is identified, it is suggested to:

- Redefine the inclusion criteria in the previous stage of the methodology to include all relevant individuals available in the database population that represent the source population.
- Generate synthetic data based on the information from other sources to cover the excluded individuals that caused exclusion bias.
- Include all relevant individuals without doing sampling in addition to the necessary inclusion and exclusion criteria from the database population. Random sampling does not cause selection bias, but other sampling methods could. Note that we are employing data mining techniques, and it is preferred to include all possible records without even random sampling from them.
- Use previously proven matching criteria in selecting controls that reduce the risk of selection bias, if the intention is to match the controls with cases.

- If the cases and controls are selected from different databases, we need to know about the characteristics of the population in those databases. If there are integrity issues in any of the databases and there is a risk of inclusion bias, it is, unfortunately, very hard to mitigate this risk, especially when there are no other appropriate database alternatives.

The level of risk on each source of bias discussed above should be reported along with the study results, especially if the mentioned suggestions are not applicable and the risks are high.

5.2.3 Step 3. Diagnostic Algorithms and Data Completeness

Risk

- Selection bias: Diagnostic bias, prevalence-incidence (Neyman) bias
- Information bias: Diagnostic bias, prevalence-incidence (Neyman) bias

Cause

Typically, cases are identified using a diagnostic algorithm (i.e., combination of various features which accurately identify people within a certain diagnosis in the datasets) that is specific — or close — to the source population's health care system. In addition, there could be comorbidity conditions as input variables that need to be constructed based on diagnostic algorithms, for example, whether the cases and controls have been previously diagnosed with another disease.

Diagnostic algorithms define a set of rules to identify whether a patient has really developed a disease or not. These rules could include conditions such as being diagnosed, whether by a general doctor or a specialist, the number of times the diagnosis is confirmed, and what

specific diagnostic codes have been recorded. These rules intend to increase the precision of identifying the actual cases. Not all diagnostic algorithms are reliable.

Depending on how and where the diagnostic algorithm is used in forming the study population and creating the study dataset, various types of bias could be introduced. Obviously, recording a case as control or vice versa due to using an inaccurate diagnostic algorithm causes diagnostic bias. This also applies to input variables which contain wrong information about the development of a disease as a matter of using the inaccurate diagnostic algorithm. Even if the diagnostic algorithm is accurate, the available data might not be complete enough to cover the disease's various types and even requested follow-up period or diagnosis repetitions in the algorithm due to sudden death, migration, and limited access to health providers of the included patients. This issue again leads to diagnostic bias.

Furthermore, if the cases are being selected using a highly inaccurate diagnostic algorithm, or the data are not complete retrospectively, that leads to identifying the prevalent cases (e.g., patients who had regular follow-up with specialists), which are a subset of the incident cases [104], this causes the prevalence-incidence or Neyman selection bias.

Actions

Identify Risk

Review the data creation plan to mark the steps in which cases are defined or comorbidity variables specified for construction, and verify the following items:

- The utilized diagnostic algorithm must be reviewed and validated by the domain experts. If an unvalidated diagnostic algorithm is used to identify the cases, there is a

risk of diagnostic selection bias, and if it is used to construct a comorbidity input variable, there is a chance of diagnostic information bias in that variable.

- There should be enough data to look back each individual's history to identify whether they were diagnosed or not. Even if the data are population-based, there could be patients that did not have follow-up for a long time. If it is not possible to look back enough, there is a risk of selection bias (in identifying the cases) and misclassification bias (in constructing comorbidity variables).

Mitigate Risk

If a risk has been identified, it is suggested to:

- Use a validated diagnostic algorithm.
    - If a validated algorithm for the focused health system is not available, a validated algorithm for another geographical region with a similar health system is sometimes acceptable.
    - If no validated algorithm is available, embark on a validation study by comparing the cases identified in the data to a reference standard group of known patients with the disease. Alternatively, the algorithm could be verified by a domain expert for face validity, however this may result in error due to the subjective nature of expert opinion.
- Select databases that include the incident cases of the study population.
- Increase the follow-up period to include all actual cases from the database population.
- Select databases that provide the ability to extend the follow-up period.
- Update the scope of study if the available cases do not represent the actual cases by clarifying that prevalent cases form the study population.

The level of risk on each source of bias discussed above should be reported along with the study results, especially if the mentioned suggestions are not applicable and the risks are high.

5.2.4 Step 4. Use of Diagnostic and Procedural Codes

Risk

- Information bias: Misclassification bias

Cause

It is a common practice to record diagnoses and procedures using codes in a health administrative system. A systematic review of administrative database research shows that exposure and outcome of patients are defined using diagnostic and procedural codes in 76% of administrative data studies [107].

Diagnostic or procedural codes recorded by health providers might not always be very accurate. The accuracy of codes is mainly affected by two factors. First, the validity of the diagnosis or procedure may differ depending on the physician (e.g., specialist vs. family physician) and locale (e.g., tertiary referral center vs. community hospital). In addition, financial incentives for physicians to use certain claims codes, or for hospitals to use certain diagnostic codes may result in increased use of those codes. For example, a patient has a cold and the code for allergic rhinitis is recorded instead of the code for viral upper respiratory tract infections (URTI). Second, the recorded code might not exactly associate with the diagnosis or procedure, especially when there are many codes that each represent a very specific type of diagnosis or procedure [106]. Verifying the codes for each individual in administrative databases is very challenging due to the large volume of data and even not having access to the test charts.

In administrative data research, the exposure and outcome variables are often constructed using diagnostic and procedural codes. It is the analyst's responsibility to identify the cases using the codes. Therefore, an error in the captured codes will cause misclassification bias, even if a validated algorithm is used to identify the cases. The same concern is applicable to building variables for exposures.

In addition, assuming the validity of the recorded codes is high, there is still risk of misclassification due to flaws in the selected codes, such as not having correct or complete association with the variable label. For example, to build a variable that demonstrates whether the patient has been exposed to general childhood infections, a list of diagnostic codes is selected that relate to the infections. If the prepared list does not cover all available general childhood infections that exist in the database's reference table of diagnostic codes, or the wrong codes are picked, the constructed variable misrepresents its title. This causes misclassification in the exposures, even if the cases were identified using strict validation methods.

In practice, it is recommended to use a validated algorithm that includes the list of diagnostic codes and the required criteria to build a variable that demonstrates the diagnosis to a disease — either as exposure or outcome. If a validated algorithm is not available, the list of diagnostic codes (e.g., ICD-10) or procedural codes (e.g., CCI) related to the disease is typically searched for and captured. Then, the new variable is constructed by checking whether any of fetched diagnostic codes have been recorded for each patient or not. Although it might sound simple, the process of capturing the list of diagnostic codes is not easy. For example, to build a new variable that shows whether the patients have been diagnosed with pneumonia (no matter what type) we search for 'pneumonia' in the ICD-10 database. J12.* to J18.* codes represent

various types of pneumonia. In addition, there are other diagnostic codes mentioned as related codes in this database under the category of pneumonia such as P23 and J82. P23 is "congenital pneumonia" and should be included with other pneumonia codes. However, J82 is "pulmonary eosinophilia," not pneumonia, so should not be included.

Furthermore, there are specific diagnostic codes that cover multiple health conditions. For example, B49 covers both fungal pneumonia and fungal meningitis. If we are collecting all diagnostic codes related to pneumonia, we will pick these codes as well. But if it has been recorded for a patient with fungal meningitis, we would still consider him as being diagnosed with a type of pneumonia.

Not all health administrative systems use International Statistical Classification of Diseases and Related Health Problems (ICD) as their diagnostic codes. Even ICD codes have different revisions. There are various types of diagnostic and procedural codes worldwide. For example, health administration systems in Ontario use the OHIP diagnostic codes, which are more generic than ICD-10. In other words, each code covers a range of particular diseases.

In addition, it is very likely that an administrative database has its own reference tables of diagnostic and procedural codes that have been derived from a standard coding system and do not include all the available codes. In this case, the analysts may need to revise the selected codes to match the almost similar ones from the reference tables. For example, a reference table for diagnostic codes could include ICD-10 codes up to one decimal point, but the code associated with the concerned disease contains two decimal points (e.g., A01.03). If the parent code (in this case A01.0) is selected, it includes cases diagnosed with the sibling codes (e.g., A01.01) that refer to another subtype of the disease as well. Therefore, whether you select the parent code or not, you will face misclassification bias. In the former, you will pick patients

diagnosed with sibling codes; in the latter, you will miss the patients diagnosed with the concerned code.

Actions

Identify Risk

Review the data creation plan to mark the databases that include codes and also the steps in which cases are defined or comorbidity variables specified for construction, and verify the following items:

- The method utilized in recording the diagnostic and procedural codes in each database should be reviewed. This includes the source of diagnosis (GP, specialist, etc.) and procedure of capturing — and transforming, if applicable — the codes. If there are concerns about why specific codes are not used or the collection method, there could be a risk of misclassification bias.
- During data exploration, if we notice a specific procedural code that had nothing to do with the diagnostics codes recorded for the same individual, this could be an error and cause misclassification bias.
- The procedure of all constructed variables should be examined. The ones constructed using diagnostic codes that cover multiple distinct diagnoses should be marked as they might cause misclassification bias.
- If diagnostic and procedural codes are captured based on a custom list of codes (i.e., reference tables) in a database and a variable is constructed using these codes, there is a risk of misclassification bias due to error in translating the expected standard code to the recorded custom code.

Mitigate Risk

Unfortunately, there is not much that can be done as the dataset is large and usually there is no access to test charts of individuals included in health administrative data. Nevertheless, it is suggested to:

- Speak to people who recorded the codes, if possible, to better understand their procedures when there are concerns about their method.
- Use codes that are recorded by providers with higher levels of certainty and investigation, such as a specialist (vs. a family physician), a hospital (vs. a clinic), etc.
- Exclude records that contain obvious contradictions between their diagnostic and procedural codes.
- Seek a domain expert's opinion if diagnostic and procedural codes are used to identify cases or construct comorbidity variables — especially if there are codes which cover a range of diagnoses or procedures.
- Use a reliable code translation atlas when working with custom codes.
- Avoid picking the parent code when custom codes with lower precision are used. However, if most of the sibling codes are related to the concerned disease and they are included in the list of codes to construct the variable, and the domain expert believes the remaining sibling codes are also not quite common, the parent code can be used.

The level of risk on each source of bias discussed above should be reported along with the study results, especially if the mentioned suggestions are not applicable and the risks are high.

5.2.5 Step 5. Time-Varying Variables

Risk

- Selection bias: Survivor treatment selection bias, diagnostic bias
- Information bias: Diagnostic bias, immortal time bias

Cause

If one or some of the variables are of a time-varying nature, for example, smoking, so that the amount of exposed time could have a significant impact on the outcome, there are various risks of time-dependent bias in this study.

In defining the status of exposures, it is important to know that the exposure happened before the outcome. If the exposure happens after the outcome, it is considered that the exposure neither has any association with the outcome nor causes the outcome. In other words, the exposure variable is null for this patient. Therefore, the time of exposure plays a significant role.

Defining the date of onset of some diseases, like IBD, is not easy due to various reasons such as long latency period and the spontaneous remitting and relapsing nature of the disease [108]. In cases where time dependency can impact the relationship between exposure and outcome, the results can face diagnostic bias.

On the other hand, unlike some variables that can have certain exposure status (e.g., appendectomy), there are time-varying variables (e.g., smoking) that impact their relationship with the outcome based on the time since exposure [104].

Another time-varying concern is about controls that do not have an index date (i.e., date of diagnosis). As mentioned earlier, when an exposure happens after the outcome, it is disregarded. However, as the controls do not have any date of diagnosis related to the outcome, the window of exploring for the exposure is longer, and, most likely, exposures are

overrepresented in the controls [104]. In other words, the exposure will mistakenly appear protective [106]. In some other situations, the certain length of follow-up might not have been considered before checking the exposure status; therefore, the exposure did not have enough time to occur [25].

Actions

Identify Risk

There are different aspects of risk when it comes to time-varying variables. Thus, the study plan and the data creation plan need to be reviewed and the following items verified:

- If the outcome is related to a disease which usually has a long latency period or spontaneous remitting or relapsing nature, it is likely that the exposure variables contain diagnostic bias.
- If an exposure variable is constructed based on an index date (e.g., date of diagnosis) for the cases, the exposure values for the controls could face survivor treatment selection bias.
- If an exposure variable is constructed based on a particular length of follow-up, it could face immortal time bias.

Mitigate Risk

If a risk has been identified, it is suggested to:

- Include as many databases as possible, which will add more accuracy in captured diagnoses and help with going back in time to check the incident date when there is a risk of diagnostic bias in an administrative data. Normally, it is not possible to verify

the diagnoses using patient charts in health administrative data due to the high volume of data.

- Consider the first prevalent date available in the accessible databases, if identifying the incident date to reduce the diagnostic bias is not likely, even though risk still remains.
- Seek past works or an expert's opinion when it comes to using variables with a time-varying nature, such as smoking, to consider a specific duration of time in which the exposure could possibly play a significant role to mitigate the immortal time bias.
- Consider the same index date of the case for the matching control when constructing the exposure variables if the controls are selected based on matching the cases to avoid the survivor treatment selection bias. If the controls are selected using another method, then there is not much that can be done to mitigate this risk.

The level of risk on each source of bias discussed above should be reported along with the study results, especially if the mentioned suggestions are not applicable and the risks are high.

5.2.6 Step 6. Lack of Contributing Variables

Risk

- Confounding: Unmeasured, measured

Cause

Administrative databases are great sources of reliable information and usually cover a large population. However, these databases are not designed for research purposes and instead are generated in the process of running a health system. Some administrative databases mainly rely on claims and therefore capture a limited variety of information, while others may be based on clinical information (e.g., from an electronic medical record) and cover a wider range

of information, although concerns about their reliability exist. Some databases can be linked to prescription or census databases to extend the information, even though there are specific limits to these databases such as not including over-the-counter drugs or having area-level data as opposed to individual-level [104]. There are frequently important variables that are not included in administrative databases that could be necessary for most research studies, such as smoking or clinical variables. This issue will definitely cause confounding in the results coming out of health administrative data. Using data mining techniques is promising in this regard as all related variables that are available can be included in the analysis and no further risks of confounding are injected during the study design, but still, the limitations in the raw data cannot be resolved.

Actions

Identify Risk

Review the actions taken during technical preprocessing and also compare the initial list of factors collected from the literature reviews and the domain expert with the list of selected attributes, and verify the following items:

- If there are factors that do not have associated variables in the Data Selection stage due to their unavailability in the selected datasets, then there is a risk of unmeasured confounding.
- If any attributes have been excluded as a matter of implementing the technical preprocessing steps, there is a risk of measured confounding.

Note that in any study for which we are limited to use only administrative data, there is always a risk of confounding.

Mitigate Risk

If the risk of confounding is high, it is suggested to:

- Add and link to more relevant and available databases to include as many selected factors as possible to further reduce the risk of confounding.
- Use methods of addressing unmeasured confounding, such as imputation of missing variables [109] or instrumental variables [110] of imputation of missing variables based on other correlated variables.
- Use data mining techniques that have adequate performance when more attributes are included for modelling so that fewer attributes are reduced during preprocessing.
- Take note of highly correlated attributes that were removed during technical preprocessing so that if their related attributes were found to play a significant role in the final models, the omitted correlated attributes are considered again during the evaluations as they might have been the confounders.

The fact that confounding risk exists in studies limited to health administrative data should be reported along with the study results, especially if a high risk of confounding is identified, and the possible confounders based on the findings should be discussed for further investigation.

5.2.7 Step 7. Handled Missing Values

Risk

- Selection bias: Exclusion bias
- Confounding: Measured

Cause

Dealing with missing values is complex, but it is usually an issue in studies conducted on real data. As a matter of handling the missing values during the technical preprocessing, it might cause elimination of some records or even variables due to the high amount of missingness in them. The former could cause an exclusion of certain characteristics of the population, which leads to risk of selection bias, while the latter will increase the chance of confounding in the results.

Actions

Identify Risk

Review the actions taken as a matter of handling the missing values in technical preprocessing and verify the following items:

- If the records eliminated due to having high amounts of missingness are mostly from specific characteristics of the population, there is a risk of exclusion bias.
- If there are variables filtered due to having high amounts of missingness, there is a risk of confounding.

Mitigate Risk

If a risk has been identified, it is suggested to:

- Review the possible solutions to handle the missing values in those records or variables and, if possible, select an alternative method that reduces the amount of elimination.

The level of risk on each source of bias discussed above should be reported along with the study results, especially if the mentioned suggestion is not applicable and the risks are high.

5.2.8 Step 8. Partitioned, Balanced, and Resampled Data

Risk

- Selection bias: Non-random sampling bias, matching bias

Cause

To build predictive models, the dataset is usually partitioned into training and test sets, then the training set is balanced to include the same number of cases and controls, and, finally, the training set is resampled to form one or multiple validation sets.

There are multiple ways to balance the data between cases and controls to build predictive models as has been discussed in the Technical Preprocessing Guideline. Even though balancing the data is a necessary step to properly train the predictive model and to increase the combined sensitivity and specificity, it changes the frequency ratio of cases and controls in the training set compared to their ratio in both the study population and source population. If the sampling from the cases or controls to balance the data is not done properly, there will be a risk of non-random sampling bias or matching bias, depending on the balancing method.

In addition, to estimate the performance of the final model during the process of model building and then to evaluate it using unseen data, resampling methods are used to generate test and validation sets. There are various resampling methods as discussed in the technical preprocessing, such as data split (used for data partitioning as well), bootstrap, k-fold cross-validation, and repeated k-fold cross-validation. In all these methods, a sample of data is selected to form test and validation sets. It is very important that the samples are selected randomly; otherwise, there is a risk of non-random sampling bias. Having said that, the

random number generator could perform differently from one platform to the other and does not always provide complete randomness.

Using different resampling methods does impact the reliability of the model's performance but has no influence on the risk of selection bias. In other words, how the records are selected to form the test and validation sets using any of the resampling methods could introduce a risk of selection bias. In addition, the final predictive model is built using the entire training set which also includes the validation sets. Therefore, eventually no record from the training set is excluded at the time of building the final model.

Actions

Identify Risk

In study plans which include building predictive models, the technical preprocessing and modelling steps need to be carefully reviewed and the following items verified to avoid selection bias:

- A random sampling method is used to partition the data into training and test sets, balance the training set, and resample the training set to generate validation set(s).
- The platform in which the experiment is implemented provides real randomness in generating random numbers.
- If a matching method is used to partition or resample the data, its criteria does not exclude specific characteristics of the population during the sampling process so that a particular part of the population is only available in one partition.

Mitigate Risk

If a risk of selection bias is identified, it is suggested to:

- Use a method that selects records randomly to partition, balance, and resample the data.
- Switch to a platform or environment that provides actual random values.
- Redefine the matching criteria to consider the excluded characteristics when sampling the data.

The risk of selection bias can be totally minimized using the above suggestions. However, if there were still concerns about a remaining risk, it needs to be reported along with the reports.

5.2.9 Step 9. Model Results

Risk

- Selection bias: Publication bias

Cause

Typically, there is no one best technique or method that builds a suitable data mining model. Instead, various models should be built and experimented with to identify the proper one. Having said that, we should only exclude the models that perform weakly in terms of accuracy or face high risk of bias that causes invalidity. Publishing results associated with particular clusters or the results that are compatible with previous works and perhaps the researcher's hypothesis per se will expose the results and conclusions to publication bias. In our context, using data mining techniques is an advantage from the perspective that no hypothesis is assumed at the beginning by allowing the algorithms to discover the relationships and patterns from the data and presenting them as models.

Actions

Identify Risk

No matter if we build one model or multiple or we use one technique or multiple, there is always a risk of publication bias as there are experimental results in the outcome of all these experiments.

Mitigate Risk

In order to reduce the risk of publication bias, it is suggested to:

- Avoid defining hypotheses and allow the data mining techniques to discover findings.
- Build models using various algorithms to have a handful number of models in the outcome.
- Ignore only the models with low performance (i.e., low accuracy) or high risk of bias (i.e., invalid findings) as they are not reliable.
- Report all remaining results as the outcome of the study and consider them all in the conclusion, regardless of whether they are compatible with the past works or our expectations.

The level of risk on the source of bias discussed above should be reported along with the study results, especially if the mentioned suggestions are not applicable and the risks are high.

6 CASE STUDY I:

IMMUNE-MEDIATED INFLAMMATORY DISEASES

The main objective of this study was to design a reliable methodology for applying data mining techniques to health administrative data. This objective was defined to set the stage and prepare an appropriate platform to address the real-world health issue, described in section 1.1, which motivated us to conduct this study. Subsequently, the designed data mining methodology needed to be evaluated. Therefore, we implemented it on case studies relevant to the concerned health conditions — immune-mediated inflammatory diseases (IMIDs) in children. These implementations helped us enhance the details in the designed methodology and verify whether the designed methodology leads to the intended goal. In addition, we expected the results of these case studies would provide new findings on the concerned health conditions. Therefore, the secondary objective of this study was to use health administrative data routinely collected from Ontario, Canada housed at ICES to build risk prediction models on children suffering with IMIDs such as inflammatory bowel disease (IBD); asthma; type 1 diabetes (T1DM); systemic autoimmune rheumatic diseases (SARDs), which include juvenile idiopathic arthritis (JIA); and multiple sclerosis (MS). There were no predefined hypotheses in the case studies as we used data mining tools to explore and discover hidden information in the data.

In Canada, provincial health administrative data are collected prospectively on all legal residents for health system planning. The data are housed at ICES and are available for research. ICES researchers are actively engaged in conducting epidemiology and health

services research using such data for pediatric-onset diseases such as diabetes [111], asthma [112], and IBD [113]. ICES's mandate is to conduct health services research that will benefit the health of Ontarians. The discipline of health services research is aimed at improving the effectiveness and efficiency of the health system, thereby reducing long-term costs and improving outcomes. In Canada, a very small portion of the population uses the majority of the health care resources. The project defined at ICES to conduct these case studies proposed to use available data to predict who will develop IMIDs based on early life attributes. In particular, we hoped to identify new hypotheses regarding risk factors associated with the outcome and predicting children with high risk of developing the diseases. Health care providers could potentially intervene earlier to either prevent disease onset (by modifying environmental risk factors) or initiate early treatment, thereby changing the clinical course of illness, resulting in better outcomes. In addition, we hoped to identify those children who will have heavy use of the health system by identifying early life predictive factors. This allows for more targeted and efficient health system interventions, improving efficiency and reducing costs. In summary, the results from these case studies aimed to improve the quality of health services for pediatric patients by predicting at-risk individuals before the development of disease and subsequent utilization of high-cost health services.

To implement the case studies, we had the choice of using the R programming language or SAS software on secure environments at ICES. The advantages of R are its open-source model and free availability, and also in its ability to model data. There are packages available for new data mining algorithms, and it allows for easy expansion through custom functions. Its main disadvantage is in data management as it reads all data into memory. This allows faster processing, but it is not appropriate for large datasets like ours and could cause crashes. In

order to overcome R's difficulties in handling large files, there are workarounds such as storing the data in an SQLite database, using parallel processing, or cloud computing. However, none of these options were accessible in our case. On the other hand, the advantages of SAS are mostly in data management as it handles the data from the hard disk and has less chance of crashing, even if the tasks take longer, whereas its disadvantages in our case are in the data modelling parts. Writing functions requires more effort, and most importantly, most of the data mining functions are under the Enterprise Miner licence, which was not accessible to us. Therefore, we decided to do the data selection and basic preprocessing actions in SAS (with the SAS Enterprise Guide application), and then continue in RStudio with the preprocessing (from the point the dataset is partitioned as the same parameters from preprocessing the training set had to be applied to the test set), modelling, and evaluation.

In this chapter, we describe our implementation of the designed methodology to predict Ontario newborns with a high risk of developing an immune-mediated inflammatory disease later in life (by age of 10 years) and generate new hypotheses of multifactorial association rules using a case-control study. The primary outcome was the incidence of any of the following: IBD, asthma, type 1 diabetes, SARDs, and MS. Therefore, patients with any of the IMIDs were considered as cases. The reason we combined all patients is that it is possible these diseases have common root causes due their similar characteristics and dysregulations they cause in the body. The goal was to build high-quality predictive models to identify significant factors, multifactorial rules, and a model that can predict patients at risk of any IMID. The details of the process to conduct this case study by applying the stages of the reliable methodology, including preparation of the IMM dataset, characteristics of the dataset, list of variables, verification of potential risks in the data and study, modelling approaches,

performance of models, repetition of modelling, extracted findings, and evaluation of results, are presented in the following representative subsections. Note that we used health administrative data available at ICES to conduct this case study; therefore, no data collection occurs as the health administrative data have already been gathered.

6.1 Stage 1: Data Selection

In this stage, we gathered a list of risk, protective, and no effect factors associated with any of the IMIDs based on literature and domain experts' opinion. Then, we selected the required datasets available from ICES that contained data related to the collected factors. Finally, we chose attributes available in those datasets if they either directly represented a factor or a derived version of them could have represented a factor. Even though the process at this stage was very clear and straightforward, it required substantial effort to explore the literature and include as many related factors as possible, go through the large list of available datasets and their documentation, and review the attributes in each dataset to understand what information is captured in each. This process could not be accomplished without the help of domain experts' guidance on the past works involving factor collection, and the assistance of database administrators and analysts at ICES who were more familiar with the available datasets, their attributes, and the relationship between datasets.

6.1.1 Factors

Many studies in epidemiology have investigated the association between various exposures and IMIDs. Some factors were detected to increase risk, some as protective, and others had no effect. We selected six articles that aggregated these findings and reviewed past studies investigating environmental factors related to pediatric IBD [9], JIA [114], T1DM [115], asthma [116], [117], and MS [118]. These articles were mostly systematic reviews suggested

by experts in each of those fields. The domain experts also suggested some potentially relevant factors such as immigration and early life factors. Overall, 128 factors were identified in the following categories:

- Perinatal
- Household (neighborhood income, rural or urban household, etc.)
- Diet
- Infection, illness, and pathogens
- Surgical interventions and medications
- Second-hand smoke
- Domestic combustion (gas cooking, biomass, etc.)
- Inhaled chemicals
- Other exposures
- Demographics and lifestyle
- Psychological
- Immigration

Note that we did not exclude any environmental factors at this point, even though it was unlikely to find relevant information for all the factors in health administrative data. A full list of retrieved factors can be seen in Appendix I (IMIDs Early Life and Environmental Factors).

6.1.2 Datasets

We planned to run the case studies using available data at ICES. However, there are hundreds of datasets stored in various data libraries. Therefore, we explored the data dictionary that provides metadata information on these datasets to select the ones that could be relevant to

our case studies. Eventually, we selected these 88 datasets from different data libraries at ICES (described in section 6.1.2.2 and Appendix III), in alphabetical order:

- ASTHMA2016 (cumulative dataset since 1991)
- CIC
- DAD1988 … DAD2016 (29 datasets)
- NACRS2006 … NACRS2016 (11 datasets)
- NIDAY
- NIDAY_LEVEL1
- NO2
- OCCC2016 (cumulative dataset since 1991)
- ODD2016 (cumulative dataset since 1991)
- OHIP1991 … OHIP2016 (26 datasets)
- PM2.5
- PSTLYEAR_LEVEL1
- PSTLYEAR_LEVEL2
- RPDBDEMO_LEVEL2
- SDS2006 … SDS2016 (11 datasets)

As these datasets all contain real-world data, we obtained privacy assessment and ethics approvals in order to use them for these case studies, in addition to specific permissions from the related organizations such as BORN Ontario and Immigration, Refugees and Citizenship Canada (IRCC). In what follows, we provide more information on obtained approvals, the data at ICES, and cohort definitions for these case studies.

6.1.2.1 Privacy and Ethics

To run these case studies, we prepared and submitted data request protocol, privacy impact assessment, and ethics documents to University of Ottawa, Children's Hospital of Eastern Ontario (CHEO), and ICES. The privacy assessment and ethics have all been approved, and their details are below:

- University of Ottawa Social Science and Humanities REB (File No: 12-16-11)
    - Approved on December 19, 2016
    - Renewed annually until 2021
- CHEO REB: Approved on February 2, 2017 (Protocol No: 17/01PE)
- ICES Privacy Impact Assessment (TRIM Number: 2017 0901 119 000)
    - Approved on July 12, 2017
    - Amendment Approved on December 11, 2018

The original ethics approval of the university can be observed in Appendix II (University of Ottawa's REB Approval).

6.1.2.2 Data

All health records of Ontario residents are linkable between databases using a unique ICES individual identification number (IKN). These data are also linkable to newborn screening metabolite data, immigration data, socio-demographic data, and medical/clinical information. In particular, a large amount of data on each child born in Ontario after 2006 can be derived by linking pregnancy data of their mother (using the Better Outcomes Registry & Network (BORN) database) to their newborn screening metabolite data (from Newborn Screening Ontario), then to early life environmental exposures based on postal code (using the Registered Persons Database), and finally to their health outcomes (using ICES health administrative

data). These very large and rich datasets could be mined to identify new early life risk factors for morbidity and later-onset immune-mediated chronic diseases.

To implement these case studies, a comprehensive dataset merged from the existing data sources was needed. The information required to extract and link ICES datasets was provided in a data creation plan submitted to ICES. The Niday dataset (from the BORN data library) was the base dataset in these case studies. BORN-Niday (in honour of Dr. Patricia Niday who, in the 1980s, envisioned the use of this data to support better obstetrical and neonatal care) includes all hospital births from fiscal year 2006 to 2011 of women who resided in Ontario with a gestational age of 20 weeks or greater and a birth weight of 500 grams or greater. Full provincial capture was obtained in the 2010/11 fiscal year in this dataset [119]. As this dataset was the main source of data in these case studies, the study population included all Ontario newborns with available data. In other words, all individuals included in the Niday dataset form the study population of more than 800,000 newborns in Ontario from fiscal 2006/07 to 2011/12. The profiles of newborns and their mothers were then linked from this dataset to the datasets of specific data libraries located at ICES (Figure 6.1) to conduct the required investigations in these case studies. These specific data libraries were ASTHMA, CIC, DAD, MOMBABY, NACRS, OCCC, ODD, OHIP, PCCF, PM2.5/NO2, RPDB, and SDS. The descriptions of these data libraries along with preliminary information on their collection and validation procedures are provided in Appendix III (Linked Data Libraries for Case Studies). The accrual start date is April 1, 2006. Also, March 31, 2017, is the accrual end date, maximum follow-up date, and when the observation window ends.

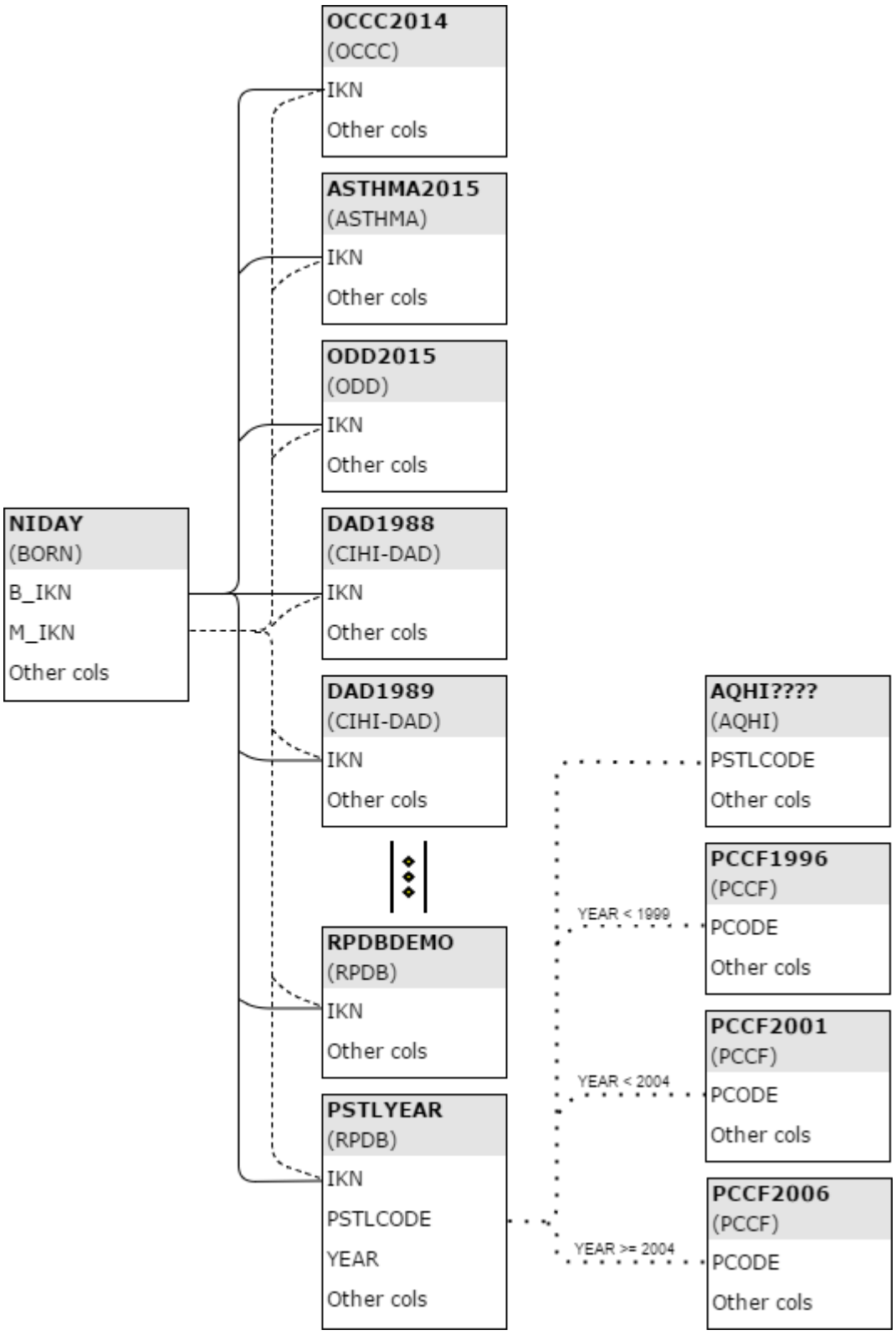

Figure 6.1: Relationships of ICES datasets to prepare the required data for case studies

6.1.2.3 Cohort Definitions

The newborns in these case studies were divided into five cohorts which demonstrate the cases. In each cohort, the incidents among newborns were needed. Note that we identified the prevalence of each disease among mothers too as part of the feature construction during the preprocessing stage.

6.1.2.3.1 IBD Cohort

Patients born in Ontario from April 1, 2006, to March 31, 2012, (FY2006–2011) contained within the Ontario Crohn's and Colitis Cohort (OCCC) dataset.

A patient aged < 18 years old is said to have an IBD and is included in OCCC if they meet one of the following criteria [120]:

- If the patient had an OHIP procedure code for sigmoidoscopy/colonoscopy, they required two hospital admissions with an IBD diagnosis or four outpatient physician OHIP claims/NACRS ER visits with an IBD diagnosis within three years, or
- If the patient did NOT have an OHIP procedure code for sigmoidoscopy/colonoscopy, they required three hospital admissions with an IBD diagnosis or seven OHIP claims/NACRS ER visits with an IBD diagnosis within three years.

Note that patients who were identified as having IBD within 6 months after birth were removed from OCCC unless they re-qualified with the child algorithm after becoming 5 years old (if so, the incident date would be the first diagnosis date after becoming 5 years old). This is based on clinical expertise (no validation due to rare cases).

6.1.2.3.2 Asthma Cohort

Patients born in Ontario from FY2006–2011 contained within the Ontario Asthma (ASTHMA) dataset.

A patient is said to be asthmatic and is included in ASTHMA if they had one hospital admission with an asthma diagnosis or two OHIP claims with an asthma diagnosis within two years [121].

6.1.2.3.3 Type 1 Diabetes Cohort

Patients born in Ontario from FY2006–2011 contained within the Ontario Diabetes Database (ODD) with incident age < 18 years old for type 1 diabetes.

A patient aged < 19 years old is said to have DM (diabetes mellitus) and is included in ODD if they had four OHIP claims with a diagnosis code of 250 or one OHIP claim with fee code of Q040, K029, K030, K045, or K046 within 2 years [122].

6.1.2.3.4 SARDs Cohort

Patients born in Ontario from FY2006–2011 contained in the DAD and OHIP databases that satisfy any of the SARD or JIA case ascertainment algorithms below.

A pediatric patient is said to have a SARD or JIA [123] if they meet the following criteria:

SARD algorithm:

One or more of the following conditions apply:

a) ≥1 hospitalization with a SARD diagnosis recorded anywhere on the hospital discharge record, or
b) ≥2 outpatient physician visits with a SARD diagnosis at least 2 months apart but within a 2-year period, or

c) ≥1 outpatient physician visit to a rheumatologist with a SARD diagnosis.

Codes included as SARDs are as following:

- ***ICD 9:*** *710.0, 710.1, 710.2, 710.3, 710.4, 710.5, 710.8, 710.9*
- ***ICD 10-CA:*** *M32.1, M32.8, M32.9, M33.x, M34.x, M35.0, M35.8, M35.9, M36.0*
- ***OHIP dxcode:*** *695, 701, 710*

JIA algorithm:

One or more of the following conditions apply:

a) ≥1 hospitalization with a JIA diagnosis recorded anywhere on the hospital discharge record, or

b) ≥2 outpatient physician visits with a JIA diagnosis at least 8 weeks apart but within a 2-year period.

Codes included as JIA are as following:

- ***ICD-9:*** *714, 720*
- ***ICD-10:*** *M05, M06, M08, M45*
- ***OHIP dxcode****: 711, 714, 715. 716*

6.1.2.3.5 MS Cohort

Patients born in Ontario from FY2006–2011 contained in DAD and OHIP databases that satisfy one or more of the following conditions [124]:

a) ≥1 hospitalization with an MS diagnosis, or

b) ≥5 physician billings coded for MS over a 2-year period.

Codes included as MS are as following:

- ***ICD-9:*** 340

- ***ICD-10:*** G35
- ***OHIP dxcode:*** 340

6.1.2.3.6 Control Cohort

The general population born in Ontario from FY2006–2011 that are not included in any of the above cohorts.

6.1.3 Attributes

After choosing the datasets, we selected all relevant attributes that could directly represent a factor or could be combined to represent a factor. Overall, 142 attributes from the selected datasets, plus 288 attributes in the PM2.5 and NO2 datasets (each attribute in these two datasets demonstrated the pollution level in every month — either in span of 1 month or 3 months — from 2005 to 2017), were selected that eventually supported 67 factors from the collected list of factors. We assigned a type to each attribute to demonstrate the role it can play in preparing the final dataset from among these options:

- **Exposure:** is representing one of more factors
- **Outcome:** is representing a target disease
- **Linkage:** is needed to link the datasets
- **Intermediary:** is needed to construct new attributes by deriving or combining their information
- **Auxiliary:** is not representing any factor but is included for any specific investigations during evaluation

Note that at this point we did not exclude any attributes due to an inaccessible privacy level or a high amount of invalid data caused by, for example, noise or missing values. These issues were considered during the preprocessing stage.

6.2 Stage 2: Preprocessing

After selecting the required data for this case study from ICES's data libraries, we had to prepare a clean dataset that was appropriate for modelling. As defined in the reliable methodology, we followed the preprocessing guidelines and applied the required actions in each step. A significant amount of time was needed to conduct the technical preprocessing. Then, we assessed the possible types of bias and confounding, and attempted to handle them as part of the contextual preprocessing. Following the designed guidelines provided peace of mind to make the best decisions in preparing the dataset and to not miss any necessary action or consideration.

6.2.1 Technical Preprocessing

We named the dataset in which IMIDs was the target variable and the children with any IMID formed its cases, the IMM dataset which initially contained 95 variables. As a result of the technical preprocessing actions, the **IMM dataset** was prepared containing **97 variables** (83 input, 1 target, 13 auxiliary variables) and **722,043 records** (split to 216,612 as test set and 505,431 as training set which was balanced to 171,898 records — 2 times 85,949 — for modelling) after preprocessing. We have summarized the technical preprocessing actions in Table 6.1. The details of each step that eventually led to the above number of variables and records are discussed in Appendix V.

As displayed in Figure 6.2, pediatrics with at least one IMID form 17% of the study population and more than 97% of them had developed asthma. Note that slightly less than a thousand cases were diagnosed with more than one IMID – therefore, the summation of the frequencies and percentages on the right chart, exceed the number of cases specified on the left chart in Figure 6.2.

Table 6.1: Technical preprocessing actions in the first case study

Step	*Title*	*Actions*
1	Dataset Creation	Base dataset: BORN-Niday After inclusion criteria: 813,719 records After exclusion criteria: 725,630 records
2	Feature Construction	Built 58 new features
3	Exploration and Planning	Init IMM dataset: 81 input, 1 target, 13 aux variables
4	Data Partitioning	70/30 data split
5	Data Cleaning	Fixed values in BWEIGHT and INTBF variables
6	Missing Values	Replaced in FSRCE Removed 5 variables: BMI, HAD_INFLUENZA, MATWGTKG, RECEIVED_ANTIVIRAL, VACCINCATION_INFLUENZA Truncated 0.5% of records [After step 9:] Imputed 3.6% of cells
7	Categorical Variables	One-hot encoded: from 76 input variables to 587
8	Feature Reduction	Removed 500 almost-unary variables Removed 4 highly correlated variables: B_SEX → SEX FSRCE.0 → M_IS_IMMIGRANT RIO2008_BIRTH → RIO2008_PREG ATOPY_IMMUNE.NA → RHINITIS_IMMUNE.NA
9	Data Normalization	Standardized all input values
10	Data Balancing	Undersampled: each class 85,949 records
11	Resampling Data	10-fold cross validation
12	Runtime Filtering	Filtered to identify important variables

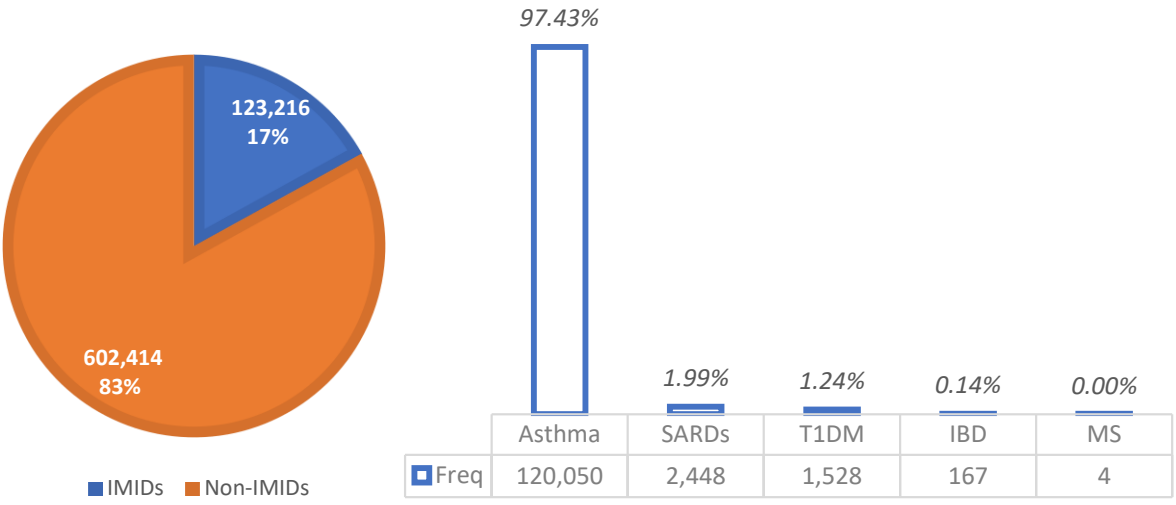

Figure 6.2: Distribution of cases vs. controls and each IMID among cases in the IMM dataset

Table 6.2: List of predictors and target variable in IMM models

Name	Label	Purpose
sex	*Sex*	Factor: sex
gest	*Gestational age at birth (weeks)*	Factor: late premature birth
deltype	*Delivery type*	Factor: c-section
intbf	*Intention to breastfeed*	Factor: breastfeeding, cow's milk formula (doesn't show mixed feeding, so if the value is TRUE the baby may still have taken formula)
labtype.(spo,ind,no)	*Labour type*	Factor: complicated delivery (receiving induction)
mathp	*Maternal health problem(s)*	Effects of maternal problems
nbres	*Newborn Resuscitation*	Effects of reviving newborn
obcomp	*Obstetrical complications*	Factor: complicated delivery
parity	*Parity*	Factor: firstborn child, household crowding, personal towel
smoking.(no,yes)	*Smoking (during preg)*	Factor: maternal smoking, passive smoking exposure

b_bmonth.(1-12)	*Baby's birth month*	Factor: fungal spore season, grass pollen exposure (Apr to Jul), sunlight exposure
m_bmonth.(1-12)	*Mother's birth month*	Effects of mother's birth month
assisted.(no,vac)	*Forceps / vacuum*	Factor: complicated delivery (assisted tools)
augment	*Augmentation*	Factor: complicated delivery (labour augmentation)
bweight	*Birth weight (grams)*	Factor: low birth weight
B_IMMUNE_INC	*Child's incidence of any immune disease*	Outcome
M_IMMUNE_PRV	*Mother's prevalence of any immune disease*	Factors related to family history
incquint_birth.(1-5)	*Nearest Census Based Neighbourhood Income Quintile at birth*	Factor: high-income lifestyle
incquint_PREG.(1-5)	*Nearest Census Based Neighbourhood Income Quintile at conception*	Factor: high-income lifestyle
rio2008_PREG	*2008 Rurality Index for Ontario at conception*	Factor: living in farm, urbanization
M_age	*Mother's age at birth*	Factor: maternal age
fsrce.(asia/pac)	*Mother's source area*	Mother/family's origin
M_IS_IMMIGRANT	*Mother is immigrant*	Factor: 2nd-gen immigrant
pm25_birth	*Air pollution level of PM2.5 in region of residence at birth*	Factor: air pollution, traffic-related particles, fine particulates
pm25_preg	*Air pollution level of PM2.5 in region of residence during pregnancy*	Factor: air pollution, traffic-related particles, fine particulates
no2_birth	*Air pollution level of NO2 in region of residence at birth*	Factor: air pollution, traffic-related particles
no2_preg	*Air pollution level of NO2 in region of residence during pregnancy*	Factor: air pollution, traffic-related particles
Gen_Child_Inf_Immune .(some ages)	*Child diagnosed with general childhood infections before incidence of first immune disease*	Factor: general childhood infection
Antibiotics_Immune .(some ages)	*Child took antibiotics before incidence of first immune disease*	Factor: use of antibiotics
Gastro_inf_Immune .(some ages)	*Child diagnosed with gastroenteritis before incidence of first immune disease*	Factor: gastroenteritis
Respir_inf_Immune .(some ages)	*Child diagnosed with respiratory infection before incidence of first immune disease*	Factor: respiratory infection

Rhinitis_Immune.na	*Child diagnosed with allergic rhinitis before incidence of first immune disease*	Factor: rhinitis
Strep_pyo_Immune.na	*Child diagnosed with Streptococcus pyogenes before incidence of first immune disease*	Factor: streptococcus pyogenes bacteria

The list of 83 predictors and one target variable used in these modelling experiments along with their descriptions are provided in Table 6.2.

6.2.2 Contextual Preprocessing

We went through the Contextual Preprocessing Guideline to investigate what systematic errors were included in this case study and whether we could have mitigated them. Note that systematic errors are separate from human random errors that can affect the results. For instance, we noticed that the values of one variable were replaced with another variable due to copying and pasting codes from one section to another. We were able to identify human random errors by analyzing the details and checking the data in various points of preprocessing and eventually reviewing the code with full attention. These types of errors are different from systematic errors, and analysts need to account for them by themselves. We have summarized the contextual preprocessing assessments that aimed to detect systematic errors in Table 6.3. The details of each step have been discussed in Appendix V.

Table 6.3: Contextual preprocessing assessments in the first case study

Step	*Title*	*Assessments*
1	Selected Source Databases	*Selection bias:* • Low risk of ascertainment bias as the sensitivity and specificity in source databases were not 100%, but still high.
2	Criteria of Study Population	*Selection bias:* • Low risk of inclusion bias as babies born out of hospitals were not included (<3% of births). • Low risk of exclusion bias as children who moved out-of-province were excluded. • No matching bias.
3	Diagnostic Algorithms and Data Completeness	*Selection bias:* • Low risk of diagnostic bias as the type of diabetes was not identified in the source data to define cases, even though validated algorithms were used for all outcomes. • No risk of prevalence-incidence bias as the incident information of cases were available. *Information bias:* • Medium risk of misclassification bias due to inaccuracy in diagnostic algorithms and underlying data to construct comorbidity variables. • Low risk of diagnostic bias as the type of diabetes was not identified in the source data to construct mother's history variables. • Low risk of prevalence-incidence bias as prevalent information of mothers was used to construct outcome history variables.
4	Use of Diagnostic and Procedural Codes	*Information bias:* • Low risk of misclassification bias due to constructing features using various diagnostic and procedural coding systems.
5	Time-Varying Variables	*Selection bias:* • Medium risk of diagnostic bias in identifying IBD patients due to short follow up period. • Low risk of survivor treatment bias in selecting controls. *Information bias:* • Low risk of immortal time bias in constructing comorbidity variables in controls.

		• High risk of misclassification bias in constructing air quality during pregnancy variables.
6	Lack of Contributing Variables	*Confounding:* • High risk of unmeasured confounding since clinical factors were not supported and dads' information was not accessible. • Low risk of measured confounding due to the elimination of mostly almost-unary variables containing limited information.
7	Handled Missing Values	*Selection bias:* • Trivial risk of exclusion bias as 0.5% of individuals were excluded. *Confounding:* • Medium risk of measured confounding due to elimination of few variables representing various factors due to high missingness.
8	Partitioned, Balanced, and Resampled Data	*Selection bias:* • Trivial risk of non-random sampling bias since random selection was used in all required sampling attempts. • No risk of matching bias as no matching method was used to do sampling.
9	Model Results	*Selection bias:* • Trivial risk of publication bias as all findings from models with reliable performance were reported.

6.3 Stage 3: Modelling

As our target variable was categorical, we built models using classification algorithms to predict whether a new patient is at high risk of developing an IMID or not, and to identify the most important variables and multifactorial rules that cause a patient to be predicted as high risk, which generates new hypotheses of factors associated to the outcome.

6.3.1 Classification

We first trained some models on a 10K subsample of our preprocessed data using multiple classifiers to check the preliminary performance of a variety of popular classifiers. Based on

their qualitative (having acceptable accuracy and not being overfit) and computational (not running very slow) performances, we selected C5 decision tree, single hidden layer neural network, and logistic regression. The classifiers considered were as follows:

- C5 (option of either rule-based or tree-based)
- C5 Tree (only decision tree)
- NNet (single hidden layer neural network)
- MLP (multilayer perceptron neural network)
- CART (classification and regression tree)
- GLM_log (logistic regression)
- SVM_poly (polynomial kernel support vector machine)
- RF (random forest)
- NB (naïve Bayes)
- KNN (k-nearest neighbors)
- Boost Tree (boosted decision tree)

C5, one of our main algorithms, produces two kinds of models: decision tree (tree-based) and rule set (rule-based). In the decision tree models, only one prediction is possible for every instance as it would belong to exactly one leaf node. In contrast, more than one rule, or no rules at all, may apply to any particular record in the rule-based model. If multiple rules apply, each gets a weighted "vote" based on the confidence associated with that rule, and the final prediction is decided by combining the weighted votes of all the rules that apply to the record in question. If no rule applies, a default prediction is assigned to the record.

It was important for us to build models that properly predict both classes (was diagnosed with IMID vs. was not diagnosed with IMID). Therefore, we selected AUC (area under the ROC

curve) as the main metric to assess the quality of the predictive models as it considers both sensitivity and specificity. Furthermore, when the models were tuned and cross-validated, their AUCs were measured and considered to select the best option. In addition to AUC, we were more interested in the sensitivity of models as that demonstrated how well the cases were being predicted and identified.

As a side experiment and in order to verify that the approach we took in undersampling the dataset to balance the classes was the right choice, we built a logistic regression model on the imbalanced version of the IMM training set. The results were poor, which assured us that balancing the data was a correct decision to provide predictive models with acceptable quality.

We built predictive models using the selected classifiers on the preprocessed (including data balancing) IMM training set. The GLM_log performed slightly worse than the other two models. C5 Tree and NNet did not have significant differences from each other. We extracted multifactorial rules from the C5 decision tree.

The attributes were ranked as each predictive model was trained. We only had access to the importance score of the top 20 variables in each model due to the limitation in the selected R package, so we calculated the adjusted score of the variables to identify those that were important across all models. Then, we rebuilt the same classifiers using the selected important variables to evaluate the results.

6.3.2 Clustering

We also attempted to identify the distinct groups of children within the study population with the goal of verifying the important variables and identifying their effect — that is, risk or protective. We clustered the preprocessed dataset using agglomerative hierarchical clustering

to suggest the number of clusters, and then built k-means partitioned clusters using the suggested number of clusters. We planned to detect the dominant value(s) for each variable within each cluster by observing the frequency of values in order to describe the main characteristics of the clusters and verify whether the identified important variables were contributing to differentiating the clusters. We did not include the target variable as part of the input variables in clustering; however, we planned to check which outcome status is dominant in each cluster to understand whether that cluster is formed mainly by cases or controls. This way, the important variables would be interpreted to have risk or protective effect. However, the k-means clustering models had very low quality, so, because the goals of these attempts were not part of our original objectives, we aborted the descriptive modelling experiments. Further details in this regard are provided in the Evaluation stage.

The functions that computed the dissimilarity distance between data points (*dist()* and *daisy()* functions in R) could not run on larger subsamples of the preprocessed dataset, let alone the entire sample, due to "out-of-memory" errors. Eventually, we were able to build agglomerative hierarchical clustering on a subsample of 30K records of the original IMM dataset (unbalanced). We observed 3 distinct clusters in the middle levels on the dendrogram plot. However, the second cluster contained the vast majority of records and was broken down into many small branches. Therefore, we decided to experiment with building k-means clustering models with k from 2 to 7.

6.4 Stage 4: Evaluation

Below, we report the results of modelling experiments and discuss the quality of models. Then we present the findings from models with satisfactory quality and discuss their validity.

6.4.1 Models

In this section, we report and discuss the results of models in the same sequence as the experiments and as described at the Modelling stage.

6.4.1.1 Classification

In the predictive models, we expected the AUC and sensitivity on unseen data to not drop significantly compared to the results from model training. If it did, then the model was overfit and not usable. Also, we expected the model's sensitivity to be acceptable so that we could rely on it. As our study is in the health context, we prefer it if more controls are predicted as cases rather than more cases being predicted as controls. The latter is dangerous, and we wanted to avoid it as much as possible.

Initially, we built some preliminary predictive models using various classifiers to select a few that perform reasonably on the IMM dataset. Table 6.4 demonstrates the results of these models.

Table 6.4: Results of preliminary classification models using 10K subsample of IMM dataset

Classifier	Model			Test			Time	Best Tune	Selected
	AUC	Sens	Spec	AUC	Sens	Spec			
C5	0.71	0.65	0.66	0.66	0.67	0.65	Medium	winnow=F, trials=10, model=rules	
C5 Tree	0.71	0.64	0.67	0.64	0.7	0.59	Medium	winnow=F, trials=10	Yes
NNet	0.72	0.67	0.65	0.66	0.75	0.58	Medium	size=1, decay=0.1	Yes
MLP	0.71	0.66	0.66	0.67	0.67	0.64	Slow	size=1	
CART	0.63	0.67	0.56	0.63	0.66	0.59	Fast	cp=0.0192	
GLM_log	0.74	0.7	0.66	0.67	0.71	0.64	Fast		Yes
SVM_poly	err	err	err	err	err	err			
RF	0.72	0.71	0.6	0.65	0.73	0.58	Slow	mtry=2	
NB	0.69	0.78	0.49	0.65	0.81	0.5	Medium	usekernel=F, fL=0	
KNN	0.56	0.66	0.51	0.56	0.62	0.5	Fast	k=7	
BoostTree	0.72	0.31	0.37	0.66	0.31	0.37	Very Slow	maxdepth=3, iter=150, nu=0.1	

There was no obligation or restriction in selecting the classifiers and we could have used all above classifiers on the entire IMM dataset. However, we preferred to take a subsample of the dataset and build various preliminary classification models to get a hint on which classifiers might perform better and faster. Our evaluation on the achieved results were as following:

- C5 Tree had almost the same AUC as the C5 (with rule-based model). We preferred to have a decision tree in order to extract strict rules from it. As mentioned earlier, C5 rules might not be able to predict some records and assign them to a class, in which case it would assign a default label or consider multiple labels for each record, which was not of interest in this experiment. So, we picked C5 Tree, especially as the sensitivity of the tree was slightly higher. Also, CART had lower performance, and Boosted Tree was very slow with poor performance.
- Among the neural networks, we decided to leave out MLP as it was slow on our data and had almost the same AUC and lower sensitivity compared to the single hidden layer neural networks. Despite that the MLP algorithm was implemented in a different R library than NNet, it can handle multiple hidden layers.
- Logistic regression gave reasonable performance and was fast. So, we picked it.
- SVM classifiers had issues with our data and could not return results, at least using the versions of algorithms we had access to in RStudio.
- Random forest was slow, while not providing outstanding results.
- Even though naïve Bayes provided quite high sensitivity, the specificity was very low, which reduced both the accuracy and AUC significantly. Still, it was an alternative due to its high sensitivity.
- KNN was interestingly fast but had quite low performance on quality.

Eventually, we decided to select the C5 Tree, NNet, and GLM_log classifiers for our main experiments on the entire IMM dataset.

Note that we built a logistic regression model on the imbalanced version of the IMM training set with over 505K records as a side experiment to verify our decision regarding data balancing at the Preprocessing stage. The majority of cases were predicted incorrectly as the sensitivity of the built model was 0.08%, while the specificity was 99% with the AUC of 53%. The detailed results were as follows:

- Trained: AUC: 0.5360 | Sens: 0.0814 | Spec: 0.9906
- Tested: AUC: 0.5356 | Sens: 0.0806 | Spec: 0.9906 | Accuracy: 0.8358
    - Confusion matrix: TP: 2,996; FN: 33,865; TN: 178,084; FP: 1,694

As can be seen, there was a very high false negative rate. These results reinforced that balancing the training set before building a classification model was a vital action. Most interestingly, the correctness accuracy of the model on the unseen data was 84%, which would cause one to think it was a high-quality model, but that result was deceiving as just the majority class (controls) was predicted very well. On the other hand, the AUC demonstrated that the model performed poorly. Even though the positive class formed 17% of the dataset, only 2% of the instances were predicted as cases. This was a sign that AUC was a more reliable metric in evaluating the quality of predictive models.

We built predictive models using the selected classifiers on the IMM dataset. The details and results of these models are reported in Table 6.5.

Table 6.5: Details and results of predictive modelling on the IMM dataset

Goal		Predict IMID using C5 Tree		Predict IMID using default NNet		Predict IMID using binomial GLM	
Run	System	*type*	ICES Intranet: R job	*type*	ICES Intranet: R job	*type*	ICES Intranet: R job
	Duration	*hours*	88.6	*hours*	88.6	*hours*	86.7
	Costly	*impute training*	59h 16m	*impute training*	59h 16m	*impute training*	59h 16m
		prep test	27h 7m	*prep test*	27h 7m	*prep test*	27h 7m
		fitting	2h	*fitting*	1h 57m	*fitting*	8m
Training Set	Imbalance	*class-1 [org:17%]*	85,949	*class-1 [org:17%]*	85,949	*class-1 [org:17%]*	85,949
		class-0	419,482	*class-0*	419,482	*class-0*	419,482
	Balance	Undersampled		Undersampled		Undersampled	
	Target	*class-1*	85,949	*class-1*	85,949	*class-1*	85,949
		class-0	85,949	*class-0*	85,949	*class-0*	85,949
Test Set		Preprocessed		Preprocessed		Preprocessed	
		Kept imbalanced		Kept imbalanced		Kept imbalanced	
Modelling	Classifier		C5		Neural Network		GLM
	Resampling		CV-10		CV-10		CV-10
	Metric		ROC		ROC		ROC
	Tuning Parameters	*model*	tree	*size*	1, 3, 5		
		winnow	FALSE, TRUE	*decay*	0, 1e-4, 0.1		
		trials	1, 10, 20				
	Samples		171,898		171,898		171,898
	Predictors		83		83		83
	Classes		2		2		2
Best Model Tune	Parameters	*model*	tree	*size*	5		
		winnow	FALSE	*decay*	0		
		trials	20				
	Specific Results	*# boosting iterations*	20	*# iterations*	100	*degrees of freedom*	171897
		avg tree size	185.9	*network*	83-5-1	*AIC*	205400
		# leaves in first tree (no boosting)	392	*weights*	426	*# fisher iteration*	4

	AUC		**0.7486**		**0.7491**		**0.7462**
	Sens		**0.6845**		**0.6853**		**0.7168**
	Spec		0.6852		0.6831		0.6454
Performance on Unseen	AUC (fitting)		0.6775		0.6805		0.6811
	Confusion Matrix	*1 -> 1 [exp: 17%]*	25,215	*1 -> 1 [exp: 17%]*	25,440	*1 -> 1 [exp: 17%]*	26,293
		1 -> 0	11,619	*1 -> 0*	11,394	*1 -> 0*	10,541
		0 -> 1	57,786	*0 -> 1*	59,371	*0 -> 1*	64,108
		0 -> 0 [exp: 83%]	121,992	*0 -> 0 [exp: 83%]*	120,407	*0 -> 0 [exp: 83%]*	115,670
	Acc		0.6796		0.6733		0.6554
	AUC		**0.6816**		**0.6802**		**0.6786**
	Sens		**0.6846**		**0.6907**		**0.7138**
	Spec		0.6786		0.6698		0.6434

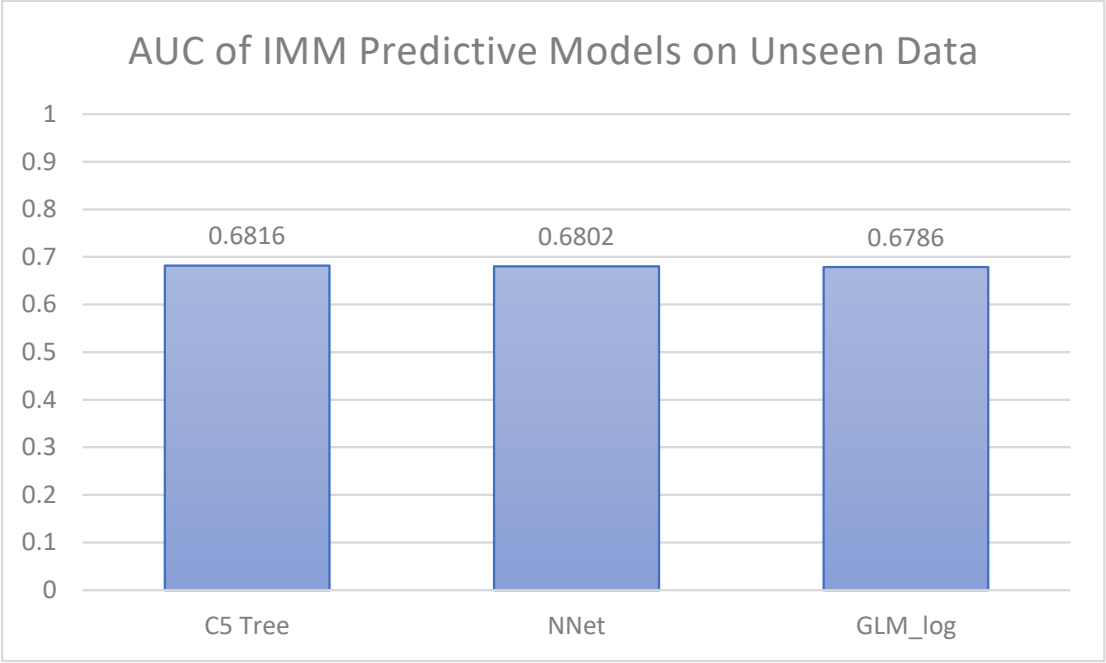

Figure 6.3: Performance of IMM predictive models on unseen data

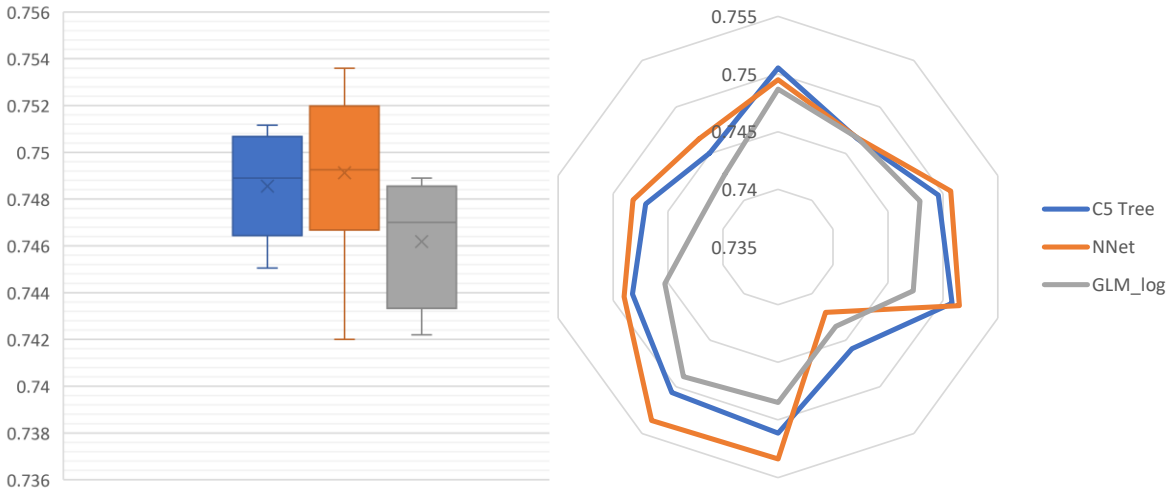

Figure 6.4: Performance of IMM predictive models during 10-fold cross validation

Considering that both the accuracy and AUC of all three models on the training and test sets was around 68%, we interpreted the results as follows:

- The sensitivity and specificity of all models were also around 68%. So, the models were predicting the cases and controls with equal performance.
- The models were not overfit and performed similarly on unseen data as on the training set that its AUC is shown in Figure 6.3. This was in favor of the models' usability.
- The AUC of 68% is notable. In other words, we could extract some interesting information, but the models could not be used for real-world prediction solutions as they did not have acceptable performance.

We ran the test of significance between the three models to identify which model performed better. The difference between the C5 Tree and the neural networks was not statistically significant ($d = 0.0006 \pm 2.26 \times 0.0006$; CV-10, CL=0.95) since the confidence interval did not span the value zero at a 95% confidence level in estimating the performance using a

10-fold cross-validation approach. However, the logistic regression had a significant difference from both the C5 Tree ($d = 0.002 \pm 2.26 \times 0.0004$; CV-10, CL=0.95) and the neural networks ($d = 0.003 \pm 2.26 \times 0.0008$; CV-10, CL=0.95), having a slightly lower performance. Their difference in performance can also be seen in Figure 6.4.

The adjusted importance scores of the variables were calculated to identify the important variables across all three models. Table 6.6 shows the list of the top 20 important variables in predicting children with high risk of developing any IMID. Note that the 11 variables in green were observed among the top 20 important variables in all models. The label and description of these variables are mentioned in Table 6.2 and Appendix IV, and we also elaborate them when evaluating the findings in section 6.4.2.3.

After the attributes were ranked and the weighted average importance score of each was calculated, the top 20 variables with the highest scores were picked and others were filtered in order to rebuild the models using the three selected classifiers. The quality of all three did not change significantly, dropping only slightly; the difference was acceptable. In other words, if we have the information from just these 20 variables, we can predict the IMID status of children with almost the same accuracy as when we had 83 input variables. The performance of the new models on unseen data was as follows:

- C5 Tree: AUC: 0.6598 | Sens: 0.6585 | Spec: 0.661
- NNet: AUC: 0.6608 | Sens: 0.6457 | Spec: 0.6759
- GLM_log: AUC: 0.659 | Sens: 0.6778 | Spec: 0.6403

Note that we also rebuilt the three selected classifiers by filtering all variables except the 11 common important variables across the main predictive models; however, the performance of the models dropped by 5%.

Table 6.6: Top 20 important variables in the first case study

Variable	*Sum*	*App*	*Avg*	*Weighted Avg*
gen_child_inf.nb	259.22	3	86.41	149.66
strep_pyo.na	253.32	3	84.44	146.25
gen_child_inf.1-3m	249.47	3	83.16	144.03
antibiotics.na	223.06	3	74.35	128.78
gen_child_inf.1	214.13	3	71.38	123.63
respir_inf.na	206.26	3	68.75	119.08
m_immune_prv	197.51	3	65.84	114.03
gastro_inf.na	187.84	3	62.61	108.45
gen_child_inf.3-12m	150.41	2	75.21	106.36
sex	174.53	3	58.18	100.76
no2_preg	96.68	1	96.68	96.68
gen_child_inf.2	167.07	3	55.69	96.46
gen_child_inf.na	95.38	1	95.38	95.38
respir_inf.1	163.23	3	54.41	94.24
parity	130.86	2	65.43	92.53
bweight	91.15	1	91.15	91.15
pm25_preg	90.86	1	90.86	90.86
respir_inf.2	127.74	2	63.87	90.33
gest	126.13	2	63.07	89.19
rio2008_preg	124.84	2	62.42	88.28

6.4.1.2 Clustering

Much of the computational time in building descriptive models was due to calculating distances between observations. The dissimilarity distance functions (*dist()* to calculate distances for hierarchical clustering and *daisy()* to calculate distances for measuring the silhouette width to verify the quality of clusters) took a significant amount of time to run, while the *kmeans()* function was quite fast and the *hclust()* was in the middle. In addition, the calculation of dissimilarity distances required an extensive amount of hardware resources

which we did not have available on the secured environment machines at ICES. Therefore, we randomly subsampled the imbalanced IMM dataset to run descriptive modelling.

The agglomerative hierarchical clustering considered each record as a separate cluster and merged the ones with the least dissimilarity distance until all clusters are merged into one — that is, a bottom-up approach. Figure 6.5 displays the dendrogram of this hierarchical clustering. Note that level 0 to 10 of the dendrogram cannot be much visible as 30K lines that each represent a record in the dataset start merging with each other from bottom to top. We observed two small clusters on the far left and far right of the dendrogram separated from other branches in the higher levels. However, the middle section was divided in many branches at the higher level. In other words, the data were not grouped into a few distinct clusters with quite obvious differentiations. Therefore, we decided not to suggest a specific number of clusters based on the built hierarchical clustering.

As we were not able to interpret a specific number of clusters from the hierarchical clustering experiment, we modelled the same subsample of the IMM dataset using k-means clustering from *k* of 2 to 7. This approach allowed us to monitor the trend in the quality of the clustering models. We measured the quality of the clusters using the silhouette width. The average silhouette width of these models is reported in Table 6.7.

Table 6.7: Performance of k-means clustering on a 30K subsample of the IMM dataset

k	*2*	*3*	***4***	*5*	*6*	*7*
Avg Sil Width	0.08	0.12	**0.13**	0.11	0.1	0.1

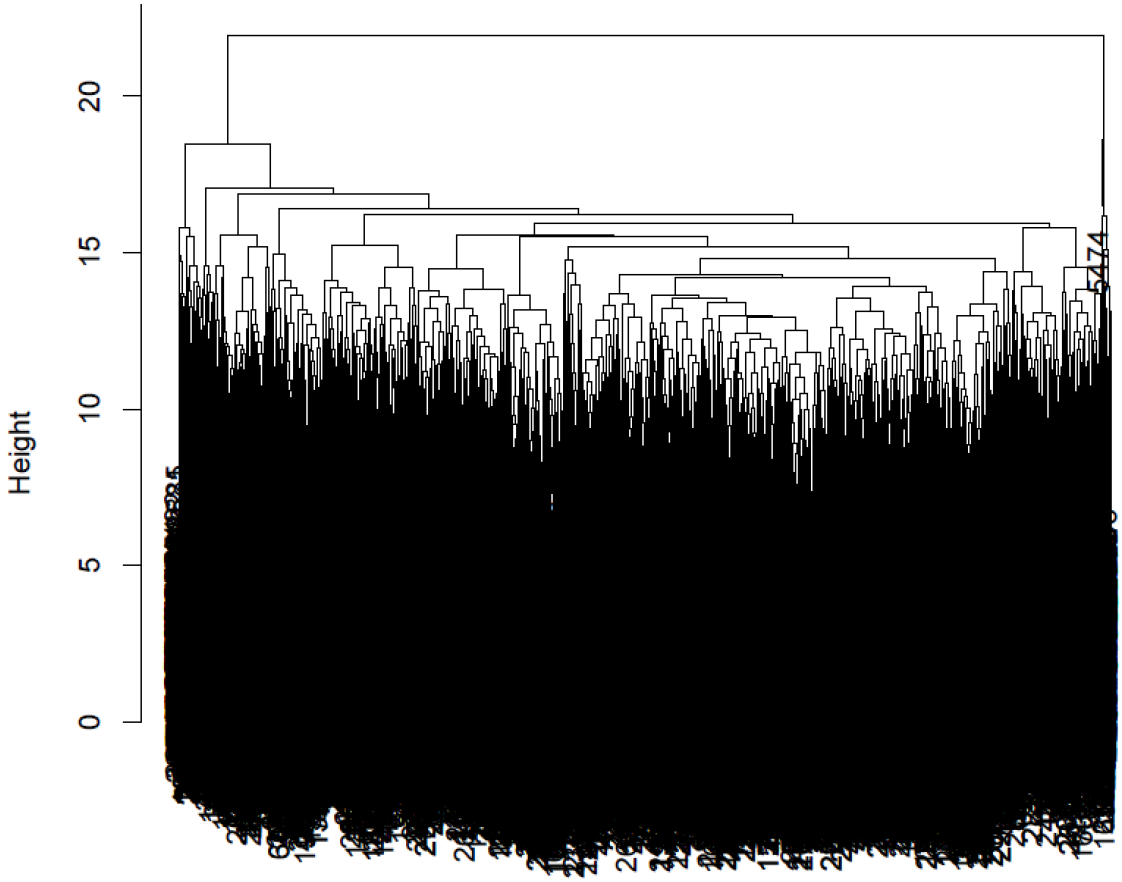

Figure 6.5: Dendrogram of agglomerative hierarchical clustering on a 30K subsample of the IMM dataset

Silhouette width is a measurement between -1 and 1 which evaluates the quality of each cluster. The larger the silhouette width is, the more cohesive the instances within the clusters and the more separated from other clusters they are. Therefore, a clustering model has high quality if the average of its clusters' silhouette widths is large and close to 1. The scores of these k-means clustering models were low; therefore, the models did not have high quality. Nevertheless, the clustering model with *k* of 4 had relatively better quality as it had a higher score. Figure 6.6 shows the silhouette width of all 4 clusters in this model.

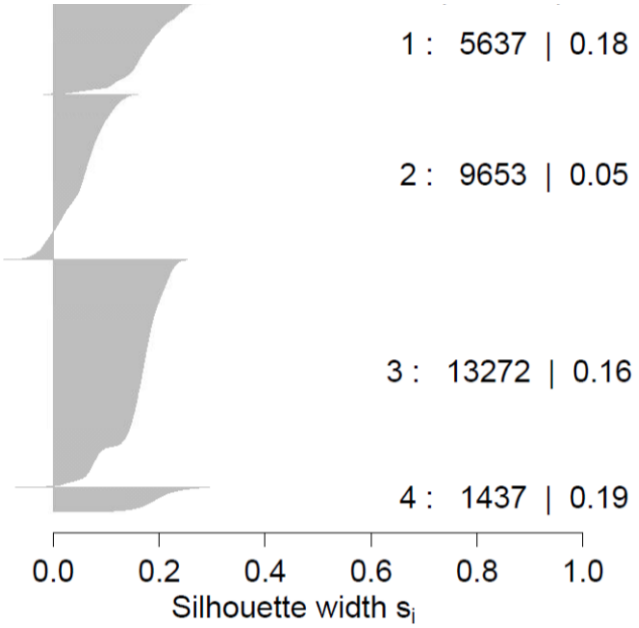

Figure 6.6: Silhouette width of clusters in k-means clustering (k=4) on a 30K subsample of the IMM dataset

The average score was below 0.2, which is considered poor quality. The second cluster had especially low quality. Therefore, we did not investigate the distribution of values in each input variable per cluster to interpret the main characteristics of each cluster. Nevertheless, Table 6.8 demonstrates the statistics of the clusters and Figure 6.7 displays the distribution of cases and controls in the clusters in the k-means clustering model with *k* of 4 on the 30K subsample of the IMM dataset. Interestingly, clusters 1 and 4 that had higher silhouette widths are mostly formed by the controls, whereas the distribution of cases and controls in the other two clusters that had lower quality are almost the same as in the entire dataset. It seems being diagnosed with IMID or not could have impact in differentiating the groups of individuals, but this model still could not distinguish between potential distinct groups of individuals existing in the data.

Table 6.8: Statistics of clusters in k-means clustering (k=4) on a 30K subsample of the IMM dataset

Label	*Cluster Size*	*Silhouette Width*	*Cases (17%)*	*Controls (83%)*
cluster-1	5,637	0.1775	598 (11%)	5,039 (89%)
cluster-2	9,653	0.0509	2,075 (21%)	7,578 (79%)
cluster-3	13,272	0.1571	2,323 (18%)	10,949 (82%)
cluster-4	1,437	0.1918	109 (8%)	1,328 (92%)

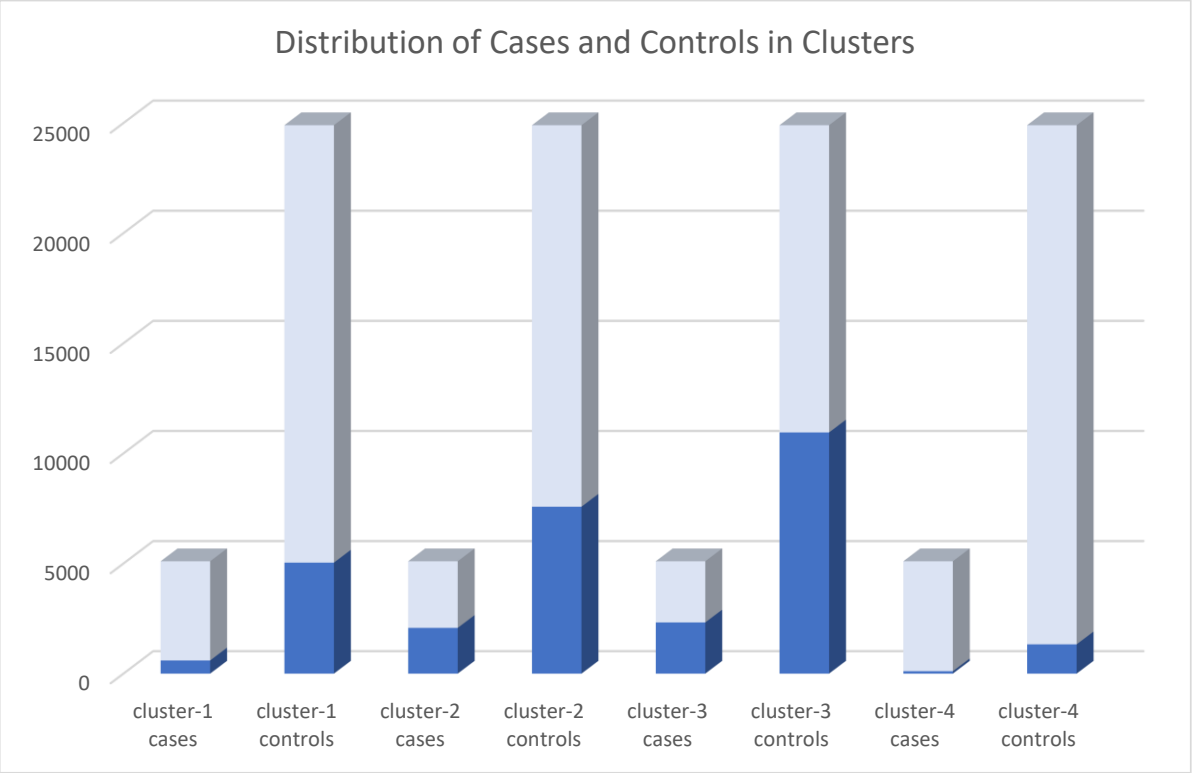

Figure 6.7: Distribution of cases and controls in clusters of k-means clustering (k=4) model

6.4.2 Findings

The findings from models with satisfactory quality are presented and discussed in this section.

6.4.2.1 Considerations

In addition to the quality of the models, which we evaluated in the previous section, the level of systematic error plays a significant role in having reliable findings. Table 6.3 summarized the level of risks in this regard. We tolerated the trivial risks, but there were still some higher risks of bias and confounding which we needed to take into consideration. To recap, we need to consider the following risks and limitations along with results and findings of this case study:

- **Confounding:** The results come with high risk of unmeasured confounding and medium risk of measured confounding as we could not include clinical factors and fathers' information. We also had to eliminate a few factors due to high missing data and many factors with almost-unary values.
- **Information bias:** The results come with high risk of misclassification bias and low risk of diagnostic bias, prevalence-incidence bias, and immortal time bias as there were inaccuracy in mothers' postal codes at conception time to construct location-based variables, significant missing values in air quality raw data which affects calculating the average value, and inaccuracy in diagnostic algorithms and underlying data to construct comorbidity variables; we did not have the diabetes type in the source data for mothers' histories and also had to use prevalent information of mothers to construct outcome history variables, various diagnostic and procedural coding systems to construct comorbidity variables, and different index dates to construct comorbidity variables in controls.

- **Selection bias:** The results come with medium risk of diagnostic bias and low risk of inclusion bias, exclusion bias, and survivor treatment bias as we had a short follow-up period in identifying IBD patients, did not have the diabetes type in the source data, could not include babies born out of hospitals, had to exclude children who had moved out of the province, and did not have a significantly long follow-up window in selecting controls.

6.4.2.2 Prediction Model

One of the findings of this case study was the predictive model itself as it could be used in production for real-world prediction solutions such as an alerting system which detects high-risk patients. The models we built were not overfit and produced proper predictions. However, their AUCs were 68%, which is considered as a notable quality rather than acceptable or high. Therefore, the models could still be used to extract some new information, but they could not be used in production.

6.4.2.3 Significant Factors

Another finding of this case study was the important variables that were among the top 20 of all three models. These variables had the highest association in predicting IMID children:

- General childhood infections, especially with first time at newborn, 1–3 months, and 1 and 2 years of age
- Streptococcus pyogenes
- Use of antibiotics
- Respiratory infection
- Mother's prevalence of any IMID
- Gastroenteritis

- Baby's sex

All these factors were among the 128 factors that had been observed in the past epidemiological studies with different effects on the IMIDs – listed in Appendix I. However, among the 67 factors that were available in our datasets, these 7 factors had the most impact in identifying IMID cases.

6.4.2.4 Multifactorial Rules

Most important were the multifactorial rules and associations we discovered from the predictive models. We extracted the paths on the decision tree that had a confidence score higher than or close to the model's performance (68%). All the extracted rules with their related details are listed in Appendix VI.

These rules were also reviewed by a domain expert and found to be interesting. Many rules included the effect of antibiotics, which, according to the content experts, was highly consistent with the literature and supported by past research. However, the rules that contained an age exclusion condition (e.g., NOT at 4 years old) caused confusion over what kind of clinical action could be derived from them. In addition, the experts seemed suspicious of why some rules contradict each other and why the ages were used in a discrete format. It is worth mentioning a few points in this regard:

1. All rules were directly discovered from the data and presented in raw format without any interpretation. In other words, there were groups of records in the dataset that comply with these rules, and we just identified them.
2. The rules that contained age exclusion conditions were described this way as they were extracted from a binary decision tree. As we had to one-hot encode the categorical variables, their presence in a decision tree would be in a format of existence or non-

existence of a categorical value when a division occurred in the tree and a path was created. One leg of each division refers to the non-existence of that categorical value as part of the conditions in that path.

3. Every rule is open to interpretation and even further investigation as a new hypothesis. For instance, when it was said “First antibiotics NOT at 4 yo” it could have meant that first antibiotics were taken before or after 4 years of age, in addition to interpreting that perhaps antibiotic use was not playing any significant role in that rule and could be ignored.
4. If one leg of the division led to a protective effect, the opposite side, when combined with other conditions, would not necessarily lead to the opposite effect. It could simply be a new path of conditions in which the first condition (i.e., the opposite leg of the concerned division) was not as important.
5. Each rule neither did cover the entire dataset nor stand alone and, instead, was relevant to the mentioned number of individuals in the dataset with the applicability score. Therefore, a more advanced interpretation could also be provided by combining relevant rules to extract information.
6. The comorbidity variables had categorical values of either not exposed to or the age at which that exposure happened for the first time. As these variables had to be one-hot encoded, a dummy variable was created for each categorical value. That is why the rules contained discrete ages. However, the combination of conditions could provide the cut-off range of age as more categorical age values that related to the same exposure were included in the associated path.
7. As most of the categorical values in the comorbidity variables had almost-unary values, they were dropped during the preprocessing. Therefore, for some of the

variables, we did not have the age-specific versions of the dummy variables; instead, just the binary presence of the exposure remained, no matter when the age of first exposure was.

In the following, we present some of the rules with higher confidence scores and notable numbers of applicable individuals. For instance, "Risk (80% on 200)" means that the focused rule applies to 200 children in our dataset in which 80% of them developed an IMID later in life. We provide a brief interpretation following each rule:

- First antibiotics at 4 yo → **Protective** (82.9% on 6,385) [IMM-1-1]
    - *Use of antibiotics for the first time in higher ages is showing a high protective effect on IMIDs. This could mean either the child was kept healthy until 4 years old and did not need antibiotics until that age, or the positive impact of antibiotics in older ages of childhood is playing a strong protective role against IMIDs.*
- First antibiotics NOT at 4 yo (could be before, after, or never), had streptococcus pyogenes → **Protective** (72.2% on 21,764) [IMM-1-2]
    - *In this rule, the usage of antibiotics does not play any role; streptococcus pyogenes is showing a fairly high protective effect on IMIDs.*
- First antibiotics at 3 yo, no streptococcus pyogenes → **Protective** (73.0% on 7,092) [IMM-1-3]
    - *Again, streptococcus pyogenes does not play any role here, and we are again observing the protective effect of using antibiotics for the first time in higher ages of childhood.*

- First antibiotics NOT at 3 or 4 yo (could be before, after, or never), no streptococcus pyogenes, first general childhood infection at 2 yo, birth weight equal or less than 2321 g, had respiratory infection, baby's birth month is April, no augmentation at birth → **Risk** (100.0% on 10) [IMM-1-5]
    - *Babies born without augmentation of labour in high allergic reactions season (and pregnancy was during seasons with less sunshine exposure) with low birth weight that developed respiratory infection are at high risk of developing IMIDs.*
- Had antibiotics but first time was NOT at 3 or 4 yo, no streptococcus pyogenes, first general childhood infection at 2 yo, birth weight greater than 2321 g → **Protective** (73.6% on 3,011) [IMM-1-6]
    - *Use of antibiotics on children with normal birth weight suggests a protective effect on IMIDs. General childhood infection is probably not playing a role as it existed in the previous risk rule too, while the change in birth weight changed the type of effect.*
- No streptococcus pyogenes, first general childhood infection NOT at 2 yo (could be before, after, or never), first antibiotics at 2 yo, gestational age greater than 32 weeks, rurality index at conception was equal or less than 12 (more urban), mom with no history of immune disease, baby boy, no smoking during pregnancy, mom is NOT an immigrant from Asia and Pacific, c-section delivery, birth month is June, PM2.5 level during pregnancy greater than 6.3 (more polluted air) → **Risk** (82.4% on 34) [IMM-1-12]

    - *Very difficult to interpret and each condition needs to be complemented with other rules that show the impact of these factors clearer. However, we can still be suspicious about c-sectioned baby boys living in urban areas with polluted air during pregnancy.*
    - *However, we should consider that the rurality index at conception has high risk of misclassification bias.*
- No streptococcus pyogenes, first antibiotics NOT at 2 or 3 or 4 yo (could be before, after, or never), had general childhood infection but first time NOT at 2 yo, first respiratory infection NOT at 2 yo (could be before, after, or never), first gastroenteritis at 2 yo, mom with no history of immune disease, rurality index at conception greater than 9 (more rural) → **Protective** (74.1% on 501) [IMM-1-16]
    - *Again, difficult to properly interpret. However, living in rural areas during pregnancy and having gastroenteritis could play a protective role.*
- No streptococcus pyogenes, no antibiotics, first general childhood infection at 1–3m of age, first respiratory infection NOT at 2 yo (could be before, after, or never), first gastroenteritis NOT at 2 yo (could be before, after, or never) → **Risk** (83.0% on 7,229) [IMM-1-20]
    - *Early general childhood infection with no use of antibiotics seems to cause high risk of IMIDs.*
- No streptococcus pyogenes, no antibiotics, first general childhood infection at 3–12m of age, first respiratory infection NOT at 2 yo (could be before, after, or never), first gastroenteritis NOT at 2 yo (could be before, after, or never), mom with history of

immune disease, gestational age greater than 34 weeks → **Risk** (81.4% on 4,017) [IMM-1-21]

- *Again, early general childhood infection with no use of antibiotics in a child whose mother has a history of IMID, even with normal gestational age, leads to high risk of developing IMIDs.*

- No streptococcus pyogenes, no antibiotics, no general childhood infection, first respiratory infection NOT at 2 yo (could be before, after, or never), first gastroenteritis NOT at 2 yo (could be before, after, or never), mom with history of immune disease, gestational age greater than 34 weeks → **Risk** (76.6% on 2,743) [IMM-1-22]
    - *This time, the mother's history of IMID in babies with normal gestational age seems to play the main role in causing high risk of developing IMIDs as the impact of other factors was eliminated.*
- No streptococcus pyogenes, first general childhood infection NOT at 2 yo (could be before, after, or never), no antibiotics, mom with history of immune disease → **Risk** (78.5% on 11,733) [IMM-4-2]
    - *This rule is from a smaller decision tree with fewer variables, and the mother's history of IMID is yet again causing high risk of developing IMIDs.*
- Had streptococcus pyogenes, mom with no history of immune disease → **Protective** (74.9% on 17,290) [IMM-6-1]
    - *We observed the protective effect of streptococcus pyogenes in an earlier rule and the negative impact of mother having history of IMID. Now, the combination of having streptococcus pyogenes and mother not having history of IMID is showing a protective effect.*

6.5 Stage 5: Feedback

To build the predictive models, we first built preliminary classification models to evaluate various classifiers. We selected three classifiers as the feedback of that attempt and returned to the Modelling stage to build the main predictive models.

The quality of the main predictive models was satisfactory, even though the performance was not at a level to become usable as standalone predictive models in production. Therefore, we proceeded to evaluate the validity of findings. We took the feedback from identifying the important variables and again returned to the Preprocessing stage to filter other variables and build new predictive models with only the selected important variables.

In the end, as the main predictive models did not have sufficient quality to become usable in production, we did not attempt to upgrade the models to a lifelong learning version.

On the other hand, we could not run clustering on the entire IMM dataset due to lack of processing resources, and the models failed. Therefore, we returned to the Preprocessing stage to take subsamples of the data and retry the modelling attempts.

The quality of clustering models was poor. Therefore, we did not proceed to evaluate the validity of findings. As the descriptive modelling attempts were not part of the study's objectives and we were experimenting with it to potentially provide added value to the predictive models, we discontinued taking more feedback and trying to improve the models. However, if we needed to do so, we would probably need to return to the Preprocessing stage and try some feature reduction options such as PCA to reduce the dimensionality. Perhaps these approaches would have helped the models to perform better and identify the more distinct clusters.

6.6 Limitations

Below, we present the limitations we faced in conducting this case study:

- **Software applications and technologies:** As discussed in the beginning of this chapter, we only had the choice of using R or SAS at ICES and could not implement the case study using other known programming languages with rich analytical packages such as Python. In addition, all implementations had to be done in a secure and protected environment as we were analyzing real-world data. Therefore, we did not have the option of using other advanced technologies in processing large amounts of data. For instance, R has difficulties in handling large files, but there are workarounds such as storing the data in a SQLite database, using parallel processing, or cloud computing which none were accessible to us. On the other hand, most of the data mining functions in SAS are under the Enterprise Miner licence, which again was not accessible to us. Therefore, we decided to use SAS for the first part of the analysis and then switch to R for the rest as earlier described in detail.
- **Missing future patients:** Our data included children born from 2006 to 2012 with a follow-up on their health status until 2017. As the outcome diseases were among chronic diseases which take time to develop, we could not include children that would be diagnosed with any of the outcomes after the follow-up date. Therefore, it is possible that some potential cases were included in the control group. For example, if a patient born in 2006 was diagnosed with IBD in 2013, then they were considered a case in our case study. However, if another patient born in the same year with similar conditions was diagnosed with IBD in 2018 (after the end of our follow-up period),

then they were among the controls in our study. This issue could have slightly distorted the models as well, but it was not possible to identify and include these cases.

- **Risk of misclassification bias in outcomes:** As discussed comprehensively in the thesis, health administrative data are highly at risk of misclassification bias in general. Specifically, the case identification in our case study were not perfect as the sensitivity of outcome-based datasets at ICES were not at 100%, and there was inaccuracy in diagnostic algorithms and captured codes. Therefore, it is likely that non-IMID patients were included in the IMIDs cases. Nevertheless, we used datasets with high sensitivity scores and validated algorithms to accomplish the best case identification possible in health administrative data and could not do anything further to overcome these limitations. In addition, the subtypes of diseases were not clearly determined in the datasets. For instance, the type of diabetes was not specified in the OHIP billing data. Therefore, we had to consider all cases in the ODD dataset diagnosed prior to 18 years of age as type 1 diabetes. It was still slightly possible that some cases diagnosed with diabetes under 18 years old were type 2 diabetes.
- **Detecting immigrant families:** The CIC dataset was used to track whether the mother of the child was an immigrant. We fetched the mother's identification key from the newborn's record and linked it to the immigration dataset to identify whether the child was from an immigrant family. However, as the father's identification key was not available, we could not identify whether the father was an immigrant. In other words, the children who had immigrant fathers but non-immigrant mothers were considered as having a non-immigrant family. To clarify this limitation, we use the "mother is immigrant" label for the related variable rather than generalizing to the whole family.

In addition, CIC data contain only information on immigrants who land in Ontario. If an immigrant arrives to another province, but subsequently moves to Ontario, they would be misclassified as a non-immigrant.

- **Small cohort sizes:** Some of the defined cohorts, such as IBD and MS, had very small sizes. In these cases, data mining techniques could not perform well in building models with higher quality and discovering more specific information.

7 CASE STUDY II: ASTHMA

To continue evaluating the reliable data mining methodology designed in this study and implementing it on case studies relevant to immune-mediated inflammatory diseases (IMIDs) in children, our second case study focused on patients diagnosed with asthma. The results from the first case study were evaluated as "notable" — meaning that the predictive models were not usable in production, even though the analysis and results were reliable. Furthermore, we still identified significant factors contributing to predicting pediatric patients with any IMID in Ontario, and we extracted interesting multifactorial rules of factors which indicate protectiveness against IMIDs or high risk of developing an IMID. However, as shown in Figure 6.2, more than 97% of pediatrics in the IMM dataset were diagnosed with asthma. Thus, we needed to clarify whether the results from the first case study were mainly relevant to asthmatic patients or if they applied to all IMIDs. Hence, we repeated the analysis, but this time only the asthmatic patients were considered as cases in order to compare the results with the first case study.

In this chapter, we describe our implementation of the designed methodology to predict Ontario newborns with a high risk of developing asthma later in life (by 10 years of age) and generate new hypotheses of multifactorial association rules using a case-control study. The primary outcome is the incidence of asthma. Therefore, only patients with asthma are considered as cases. If the results are similar to the ones from the first case study, it would mean the earlier predictive models were biased toward asthma and the extracted rules might not be very relevant to other IMIDs. The goal was to build high-quality predictive models to

identify significant factors, multifactorial rules, and a model that can predict patients at risk of asthma in order to compare with the results from the first case study. The details of the process to conduct this case study by applying the stages of the reliable methodology, including preparation of the ASTH dataset, characteristics of the dataset, list of variables, verification of potential risks in the data and study, approaches to modelling, performance of models, repetition of modelling, extracted findings, and evaluation of results, are presented in the following subsections. Note that we used health administrative data available at ICES to conduct this case study; therefore, no data collection occurred, as the health administrative data had already been gathered.

7.1 Stage 1: Data Selection

We had already gathered a list of risk, protective, and no effect factors associated with asthma based on literature and domain experts' opinion; selected the required datasets available from ICES; and chosen attributes available in those datasets related to the collected factors in the first case study.

7.1.1 Factors

In section 6.1.1, we mentioned the procedure and sources to collect the 128 factors in various categories. A full list of retrieved factors can be seen in Appendix I (IMIDs Early Life and Environmental Factors). Note that we included all collected factors in this case study, even if they were observed to be associated with other IMIDs, as it is possible these diseases have common root causes due to their similar characteristics and the dysregulations they cause in the body.

7.1.2 Datasets

We used the same selection of 88 datasets from different data libraries at ICES as described in section 6.1.2 for this case study. Even though some datasets contained information related to specific IMIDs, such as the OCCC dataset for IBD patients, we did not exclude them in this stage as they still could include variables useful for investigations on the results. Further information on the data selection procedures including the privacy assessment, ethics approvals, and cohort definitions are provided in the same section.

7.1.3 Attributes

As reported in section 6.1.3, 142 attributes from the selected datasets plus 288 attributes in the PM2.5 and NO2 datasets were selected. They eventually supported 67 factors from the collected list of factors. We assigned a type to each attribute to demonstrate the role it can play in preparing the final dataset from among these options: Exposure, Outcome, Linkage, Intermediary, and Auxiliary.

7.2 Stage 2: Preprocessing

As the prepared dataset in the first case study contained variables that were constructed based on each individual's outcome status, we could not use the IMM dataset for this case study. However, following the guidelines defined in the reliable methodology, most preprocessing actions were similar to the first case study as the same data were selected to be preprocessed, just for a different outcome. Therefore, the same information described in Appendix V (Preprocessing Details of the First Case Study) were applicable to this case study with slightly different details. The following sections summarize each preprocessing step and point out the differences with the preprocessing actions taken in the first case study.

7.2.1 Technical Preprocessing

We named the dataset in which asthma was the target variable and the children with asthma formed its cases, the ASTH dataset which initially contained 96 variables. As a result of the technical preprocessing actions, the **ASTH dataset** was prepared containing **98 variables** (84 input, 1 target, 13 auxiliary variables) and **722,038 records** (split to 216,610 as the test set and 505,428 as the training set, which was balanced to 167,478 records — 2 times 83,739 — for modelling) after preprocessing. We have summarized the technical preprocessing actions in Table 7.1. The ones different from the first case study are underlined.

The list of 84 predictors and one target variable used in these modelling experiments along with their descriptions is provided in Table 7.2.

7.2.2 Contextual Preprocessing

The potential systematic errors in conducting this case study were similar to those identified and described in the first case study, which are discussed in Appendix V. As an input variable reflected the prevalence of all IMIDs among mothers, the concerns in identifying patients with IMIDs other than asthma were applicable in information biases but not in selection biases. We have listed the contextual preprocessing assessments that aimed to detect systematic errors in Table 7.3. The ones different from those in the first case study are underlined.

7.3 Stage 3: Modelling

As our target variable was categorical, we built models using classification algorithms to predict whether a new patient is at high risk of developing asthma and to identify the most important variables and multifactorial rules that cause a patient to be predicted as high risk, which generates new hypotheses of factors associated to the outcome.

Table 7.1: Technical preprocessing actions in the second case study

Step	*Title*	*Actions*
1	Dataset Creation	Base dataset: BORN-Niday After inclusion criteria: 813,719 records After exclusion criteria: 725,630 records
2	Feature Construction	Built 58 new features
3	Exploration and Planning	Init ASTH dataset: 82 input, 1 target, 13 aux variables *{included M_IMMUNE_PRV in addition to M_ASTHMA_PRV in inputs}*
4	Data Partitioning	70/30 data split
5	Data Cleaning	Fixed values in BWEIGHT and INTBF variables
6	Missing Values	Replaced in FSRCE Removed 5 variables: BMI, HAD_INFLUENZA, MATWGTKG, RECEIVED_ANTIVIRAL, VACCINCATION_INFLUENZA Truncated 0.5% of records [After step 9:] Imputed 2.1% of cells *{this random selection of training records contained fewer missing cells}*
7	Categorical Variables	One-hot encoded: from 77 input variables to 588
8	Feature Reduction	Removed 499 almost-unary variables Removed 5 highly correlated variables: B_SEX → SEX MATHP2 → M_ASTHMA_PRV FSRCE.0 → M_IS_IMMIGRANT RIO2008_BIRTH → RIO2008_PREG ATOPY_ASTHMA.NA → RHINITIS_ASTHMA.NA
9	Data Normalization	Standardized all input values
10	Data Balancing	Undersampled: each class 83,739 records *{there are fewer cases}*
11	Resampling Data	10-fold cross validation
12	Runtime Filtering	None

Table 7.2: List of predictors and target variable in ASTH models

Name	Label	Purpose
sex	*Sex*	Factor: sex
gest	*Gestational age at birth (weeks)*	Factor: late premature birth
deltype	*Delivery type*	Factor: c-section
intbf	*Intention to breastfeed*	Factor: breastfeeding, cow's milk formula (doesn't show mixed feeding, so if the value is TRUE the baby may still have taken formula)
labtype.(spo,ind,no)	*Labour type*	Factor: complicated delivery (receiving induction)
mathp	*Maternal health problem(s)*	Effects of maternal problems
nbres	*Newborn Resuscitation*	Effects of reviving newborn
obcomp	*Obstetrical complications*	Factor: complicated delivery
parity	*Parity*	Factor: firstborn child, household crowding, personal towel
smoking.(no,yes)	*Smoking (during preg)*	Factor: maternal smoking, passive smoking exposure
b_bmonth.(1-12)	*Baby's birth month*	Factor: fungal spore season, grass pollen exposure (Apr to Jul), sunlight exposure
m_bmonth.(1-12)	*Mother's birth month*	Effects of mother's birth month
assisted.(no,vac)	*Forceps / vacuum*	Factor: complicated delivery (assisted tools)
augment	*Augmentation*	Factor: complicated delivery (labour augmentation)
bweight	*Birth weight (grams)*	Factor: low birth weight
B_ASTHMA_INC	*Child's incidence of asthma*	Outcome
M_ASTHMA_PRV	*Mother's prevalence of asthma*	Factors related to family history
M_IMMUNE_PRV	*Mother's prevalence of any immune disease*	Factors related to family history
incquint_birth.(1-5)	*Nearest Census Based Neighbourhood Income Quintile at birth*	Factor: high-income lifestyle

incquint_PREG.(1-5)	*Nearest Census Based Neighbourhood Income Quintile at conception*	Factor: high-income lifestyle
rio2008_PREG	*2008 Rurality Index for Ontario at conception*	Factor: living in farm, urbanization
M_age	*Mother's age at birth*	Factor: maternal age
fsrce.(asia/pac)	*Mother's source area*	Mother/family's origin
M_IS_IMMIGRANT	*Mother is immigrant*	Factor: 2nd-gen immigrant
pm25_birth	*Air pollution level of PM2.5 in region of residence at birth*	Factor: air pollution, traffic-related particles, fine particulates
pm25_preg	*Air pollution level of PM2.5 in region of residence during pregnancy*	Factor: air pollution, traffic-related particles, fine particulates
no2_birth	*Air pollution level of NO2 in region of residence at birth*	Factor: air pollution, traffic-related particles
no2_preg	*Air pollution level of NO2 in region of residence during pregnancy*	Factor: air pollution, traffic-related particles
Gen_Child_Inf_ASTHMA .(some ages)	*Child diagnosed with general childhood infections before incidence of asthma*	Factor: general childhood infection
Antibiotics_ASTHMA .(some ages)	*Child took antibiotics before incidence of asthma*	Factor: use of antibiotics
Gastro_inf_ASTHMA .(some ages)	*Child diagnosed with gastroenteritis before incidence of asthma*	Factor: gastroenteritis
Respir_inf_ASTHMA .(some ages)	*Child diagnosed with respiratory infection before incidence of asthma*	Factor: respiratory infection
Rhinitis_ASTHMA.na	*Child diagnosed with allergic rhinitis before incidence of asthma*	Factor: rhinitis
Strep_pyo_ASTHMA.na	*Child diagnosed with Streptococcus pyogenes before incidence of asthma*	Factor: streptococcus pyogenes bacteria

Table 7.3: Contextual preprocessing assessments in the second case study

Step	*Title*	*Assessments*
1	Selected Source Databases	*Selection bias:* • Low risk of ascertainment bias as the sensitivity and specificity in source databases were not 100%, but still high.
2	Criteria of Study Population	*Selection bias:* • Low risk of inclusion bias as babies born out of hospitals were not included (<3% of births). • Low risk of exclusion bias as children who moved out-of-province were excluded. • No matching bias.
3	Diagnostic Algorithms and Data Completeness	*Selection bias:* • No risk of diagnostic bias as diabetic patients were not part of the cases so that lack of type of diabetes in source data could cause an issue. • No risk of prevalence-incidence bias as the incident information of cases were available. *Information bias:* • Medium risk of misclassification bias due to inaccuracy in diagnostic algorithms and underlying data to construct comorbidity variables. • Low risk of diagnostic bias as the type of diabetes was not identified in the source data to construct mother's history variables. • Low risk of prevalence-incidence bias as prevalent information of mothers was used to construct outcome history variables.
4	Use of Diagnostic and Procedural Codes	*Information bias:* • Low risk of misclassification bias due to constructing features using various diagnostic and procedural coding systems.
5	Time-Varying Variables	*Selection bias:* • No risk of diagnostic bias in identifying IBD patients due to short follow up period as they are not part of cases. • Low risk of survivor treatment bias in selecting controls. *Information bias:* • Low risk of immortal time bias in constructing comorbidity variables in controls.

		• Low risk of misclassification bias in constructing location-based variables during pregnancy time.
6	Lack of Contributing Variables	*Confounding:* • High risk of unmeasured confounding since clinical factors were not supported and dads' information was not accessible. • Low risk of measured confounding due to the elimination of mostly almost-unary variables containing limited information.
7	Handled Missing Values	*Selection bias:* • Trivial risk of exclusion bias as 0.5% of individuals were excluded. *Confounding:* • Medium risk of measured confounding due to elimination of few variables representing various factors due to high missingness.
8	Partitioned, Balanced, and Resampled Data	*Selection bias:* • Trivial risk of non-random sampling bias since random selection was used in all required sampling attempts. • No risk of matching bias as no matching method was used to do sampling.
9	Model Results	*Selection bias:* • Trivial risk of publication bias as all findings from models with reliable performance were reported.

We used the same three classifiers from the first case study as various classifiers were tested in the preliminary experiment of that case study and the data between the two case studies did not have significant differences. The selected classifiers were the C5 decision tree, single hidden layer neural network, and logistic regression. In addition, we used AUC (area under the ROC curve) as the main metric to assess the quality of the predictive models throughout the entire process and considered sensitivity as well due to the same reasons explained in the first case study.

We built predictive models using the selected classifiers on the preprocessed (including data balancing) ASTH training set. The GLM_log performed slightly worse than the other two models. The C5 Tree and NNet did not have significant differences from each other. We extracted multifactorial rules from the C5 decision tree.

The attributes were ranked as each predictive model was trained. We only had access to the importance score of the top 20 variables in each model due to the limitation in the selected R package, so we calculated the adjusted score of the variables to identify those that were important across all models.

7.4 Stage 4: Evaluation

Below, we report the results of modelling experiments and discuss the quality of the models. Then, we present the findings from models with satisfactory quality and discuss their validity.

7.4.1 Models

In the predictive models, we expected the AUC and sensitivity on unseen data to not drop significantly compared to the results from the model training. If they did, then the model was overfit and not usable. Also, we expected the model's sensitivity to be acceptable so that we could rely on it. As our study is in the health context, we prefer that more controls are predicted as cases rather than more cases being predicted as controls. The latter is dangerous, and we wanted to avoid it as much as possible.

We built predictive models using the selected classifiers on the ASTH dataset. The details and results of these models are reported in Table 7.4.

Table 7.4: Details and results of predictive modelling on the ASTH dataset

Goal		Predict Asthma using C5 Tree		Predict Asthma using default Nnet		Predict Asthma using binomial GLM	
Run	System	*type*	ICES Intranet: R job	*type*	ICES Intranet: R job	*type*	ICES Intranet: R job
	Duration	*hours*	96.5	*hours*	96.7	*hours*	94.9
	Costly	*impute training*	69h 40m	*impute training*	69h 40m	*impute training*	69h 40m
		prep test	24h 47m	*prep test*	24h 47m	*prep test*	24h 47m
		fitting	1h 46m	*fitting*	1h 57m	*fitting*	8m
Training Set	Imbalance	*class-1 [org:16.6%]*	83,739	*class-1 [org:16.6%]*	83,739	*class-1 [org:16.6%]*	83,739
		class-0	421,689	*class-0*	421,689	*class-0*	421,689
	Balance	Undersampled		Undersampled		Undersampled	
	Target	*class-1*	83,739	*class-1*	83,739	*class-1*	83,739
		class-0	83,739	*class-0*	83,739	*class-0*	83,739
Test Set		Preprocessed		Preprocessed		Preprocessed	
		Kept imbalanced		Kept imbalanced		Kept imbalanced	
Modelling	Classifier		C5		Neural Network		GLM
	Resampling		CV-10		CV-10		CV-10
	Metric		ROC		ROC		ROC
	Tuning Parameters	*model*	tree	*size*	1, 3, 5		
		winnow	FALSE, TRUE	*decay*	0, 1e-4, 0.1		
		trials	1, 10, 20				
	Samples		167,478		167,478		167,478
	Predictors		84		84		84
	Classes		2		2		2
Best Model Tune	Parameters	*model*	tree	*size*	5		
		winnow	FALSE	*decay*	0.1		
		trials	20				
	Specific Results	*# boosting iterations*	20	*# iterations*	100	*degrees of freedom*	167477
		avg tree size	182.9	*network*	84-5-1	*AIC*	199700
		# leaves in first tree (no boosting)	359	*weights*	431	*# fisher iteration*	4

	AUC		**0.7510**		**0.7506**		**0.7478**
	Sens		**0.6808**		**0.6770**		**0.7139**
	Spec		0.6915		0.6927		0.6510
Performance on Unseen	AUC (fitting)		0.6788		0.6830		0.6825
	Confusion Matrix	*1 -> 1 [exp: 16.6%]*	24,325	*1 -> 1 [exp: 16.6%]*	24,212	*1 -> 1 [exp: 16.6%]*	25,617
		1 -> 0	11,562	*1 -> 0*	11,675	*1 -> 0*	10,270
		0 -> 1	55,216	*0 -> 1*	54,916	*0 -> 1*	62,949
		0 -> 0 [exp: 83%]	125,507	*0 -> 0 [exp: 83%]*	125,807	*0 -> 0 [exp: 83%]*	117,774
	Acc		0.6917		0.6926		0.6620
	AUC		**0.6861**		**0.6854**		**0.6828**
	Sens		**0.6778**		**0.6747**		**0.7138**
	Spec		0.6945		0.6961		0.6517

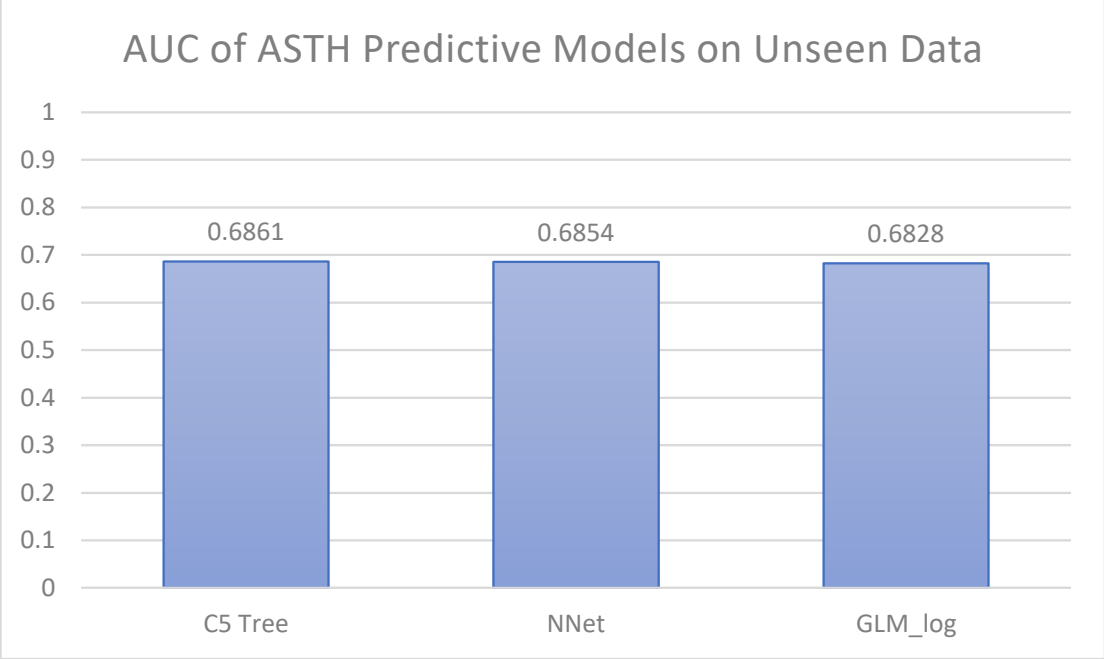

Figure 7.1: Performance of ASTH predictive models on unseen data

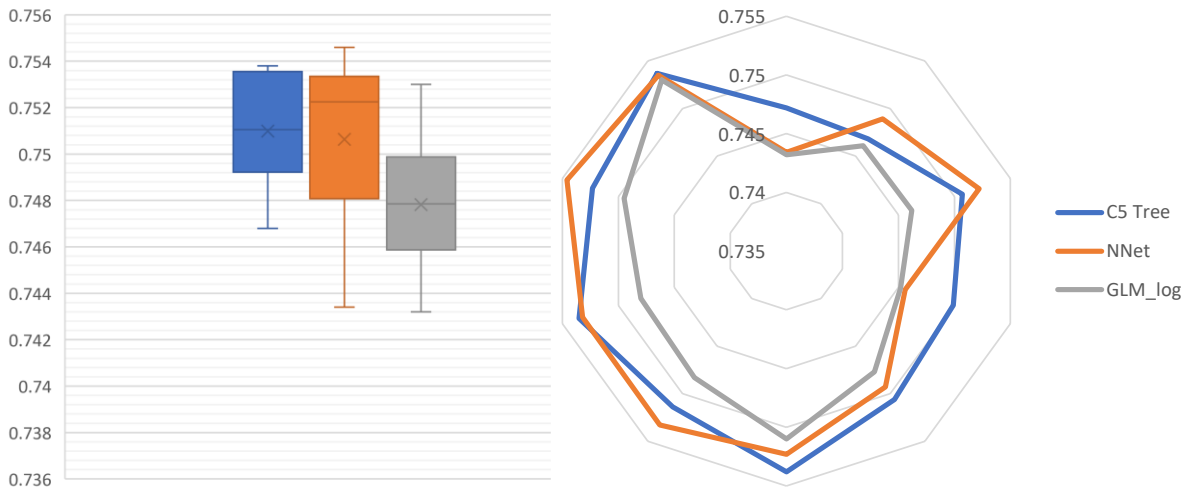

Figure 7.2: Performance of ASTH predictive models during 10-fold cross validation

Considering that both the accuracy and AUC of all three models on the training and test sets were around 69%, we interpreted the results as follows:

- The sensitivity and specificity of all models were also around 69%. So, the models predicted the cases and controls with equal performance.
- The models were not overfit and performed similarly on unseen data as on the training set. The AUCs are shown in Figure 7.1. This was in favor of the models' usability.
- The AUC of 69% is notable as it represents only weak to moderate prediction. In other words, we could extract some interesting information, but the models could not be used for real-world prediction solutions as they did not have acceptable performance.

We ran the test of significance between the three models to identify which model performed better. The difference between the C5 Tree and the neural network was not statistically significant ($d = 0.0004 \pm 2.26 \times 0.0008$; CV-10, CL=0.95) since the confidence interval did not span the value zero at a 95% confidence level in estimating the performance using a

10-fold cross-validation approach. However, the logistic regression had a significant difference from both the C5 Tree ($d = 0.003 \pm 2.26 \times 0.0005$; CV-10, CL=0.95) and the neural networks ($d = 0.003 \pm 2.26 \times 0.0007$; CV-10, CL=0.95) with a slightly lower performance. The difference in performance can also be seen in Figure 7.2.

The adjusted importance scores of the variables were calculated to identify the important variables across all three models. Table 7.5 shows the list of the top 20 important variables in

Table 7.5: Top 20 important variables in the second case study

Variable	*Sum*	*App*	*Avg*	*Weighted Avg*
strep_pyo.na	251.55	3	83.85	145.23
gen_child_inf.nb	247.46	3	82.49	142.87
antibiotics.na	242.95	3	80.98	140.27
gen_child_inf.1-3m	240.94	3	80.31	139.11
antibiotics.4	172.05	2	86.03	121.66
respir_inf.na	201.33	3	67.11	116.28
gen_child_inf.1	194.47	3	64.82	112.28
sex	154.65	2	77.33	109.35
gen_child_inf.3-12m	151.72	2	75.86	107.28
antibiotics.3	151.11	2	75.56	106.85
antibiotics.2	136.74	2	68.37	96.69
bweight	96.24	1	96.24	96.24
respir_inf.1	165.95	3	55.32	95.81
parity	133.27	2	66.64	94.24
gen_child_inf.2	160.79	3	53.6	92.83
rio2008_preg	130.89	2	65.45	92.55
m_asthma_prv	128.59	2	64.3	90.93
respir_inf.2	127.59	2	63.8	90.22
gest	127.33	2	63.67	90.04
pm25_preg	89.01	1	89.01	89.01

predicting children with high risk of developing asthma. Note that the 8 variables in green were observed among the top 20 important variables in all models. The label and description of these variable names are mentioned in Table 7.2 and Appendix IV, and we also elaborate them when evaluating the findings in section 7.4.2.3.

7.4.2 Findings

The findings from models with satisfactory quality are presented and discussed in this section. We also compare the findings of the first and second case studies. As the results and findings between the two case studies are very similar, it seems the results of predicting IMID patients were dominated by asthmatic patients. This conclusion was very likely because more than 97% of IMID cases were from the asthma cohort, as displayed in Figure 6.2. The comparison of each finding is discussed in the following sections.

7.4.2.1 Considerations

In addition to the quality of the models, which we evaluated in the previous section, the level of systematic error plays a significant role in having reliable findings. Table 7.3 summarized the level of risks in this regard. We tolerated the trivial risks, but there were still some higher risks of bias and confounding which we needed to take into consideration. To recap, we need to consider the following risks and limitations along with results and findings of this case study:

- **Confounding:** The results come with high risk of unmeasured confounding and medium risk of measured confounding as we could not include clinical factors and fathers' information. We also had to eliminate a few factors due to high missing data and many factors with almost-unary values.

- **Information bias:** The results come with high risk of misclassification bias and low risk of diagnostic bias, prevalence-incidence bias, and immortal time bias as there were inaccuracy in mothers' postal codes at conception time, which were used to construct location-based variables; significant missing values in air quality raw data, which affects calculating the average value; and inaccuracy in diagnostic algorithms and underlying data used to construct comorbidity variables. Moreover, we did not have the diabetes type in the source data for mothers' histories and also had to use prevalence information of mothers to construct outcome history variables, various diagnostic and procedural coding systems to construct comorbidity variables, and different index dates to construct comorbidity variables in controls.
- **Selection bias:** The results come with low risk of inclusion bias, exclusion bias, and survivor treatment bias as we could not include babies born out of hospitals, had to exclude children who had moved out of the province, and did not have a significantly long follow-up window in selecting controls.

7.4.2.2 Prediction Model

One of the findings of this case study was the predictive model itself as it could be used in production for real-world prediction solutions such as an alerting system which detects high-risk patients. The models we built were not overfit and produced proper predictions. However, their AUCs were 69%, which is considered weak to moderate predictive ability. Therefore, the models could still be used to extract new information, but they could not be used in production.

The performance of predictive models in both case studies was very similar. In both case studies, the logistic regression model performed slightly worse than the other two models,

which did not have significant differences between them. In addition, we ran tests of significance on models built by the same classifier between the two case studies. In other words, we compared the performance of the C5 Tree model from the IMM dataset with the C5 Tree model from the ASTH dataset, then compared the neural network models of each case study and, similarly, the logistic regression models. None of the pairs of models had significant differences. Hence, the prediction behaviour of models on IMIDs and asthma patients are the same.

7.4.2.3 Significant Factors

Another finding of this case study was the important variables that were among the top 20 of all three models. These variables had the highest association in predicting asthmatic children:

- General childhood infections, especially with first time at newborn, 1–3 months, and 1 and 2 years of age
- Streptococcus pyogenes
- Use of antibiotics
- Respiratory infection, especially with first time at 1 year of age

Comparing these factors to the significant factors from the first case study, they completely overlap with each other. The other three significant factors that ranked among the top 20 important variables of all three models in the first case study were still among the top 20 factors in the second case study, although not in all three models.

7.4.2.4 Multifactorial Rules

We extracted the paths on the decision tree that had a confidence score higher than or close to the model's performance (69%). All the extracted rules with their related details are listed in

Appendix VII. Further information on how to interpret these rules is also provided in section 6.4.2.4.

Comparing the extracted rules with higher confidence scores and notable number of applicable individuals to the ones discovered in the first case study, we see they are highly similar. Below, we list some of these multifactorial rules that are almost identical along with their identification numbers in both case studies:

- ASTH-1-1 [identical to IMM-1-1]: First antibiotics at 4 yo → **Protective**
- ASTH-1-2 [similar to IMM-1-2]: First antibiotics NOT at 4 yo (could be before, after, or never), had streptococcus pyogenes, mother with no history of asthma → **Protective**
- ASTH-1-5 [identical to IMM-1-3]: First antibiotics at 3 yo, no streptococcus pyogenes → **Protective**
- ASTH-1-7 [similar to IMM-1-5]: First antibiotics NOT at 3 or 4 yo (could be before, after, or never), no streptococcus pyogenes, first general childhood infection at 2 yo, birth weight equal or less than 2989 g, baby's birth month is NOT December, mother is an immigrant → **Risk**
- ASTH-1-6 [identical to IMM-1-6]: Had antibiotics but first time was NOT at 3 or 4 yo, no streptococcus pyogenes, first general childhood infection at 2 yo, birth weight greater than 2989 g → **Protective**
- ASTH-1-10 [similar to IMM-1-12]: No streptococcus pyogenes, first general childhood infection NOT at 2 yo (could be before, after, or never), first antibiotics at 2 yo, gestational age greater than 34 weeks, rurality index at conception was greater than 11 (more rural), mom with no history of asthma, first respiratory infection at 1–

3m of age, birth month is NOT February or September, no rhinitis, income quintile at birth was NOT 3, no labour during birth → **Risk**

    - *Note that the rurality index at conception has high risk of misclassification bias.*

- ASTH-1-45 [similar to IMM-1-21]: No streptococcus pyogenes, no antibiotics, first general childhood infection at 1–3m of age, first respiratory infection NOT at 2 yo (could be before, after, or never), first gastroenteritis NOT at 2 yo (could be before, after, or never), mom with history of asthma → **Risk**
- ASTH-1-47 [similar to IMM-1-22]: No streptococcus pyogenes, no antibiotics, first general childhood infection at 3–12m of age, first respiratory infection NOT at 2 yo (could be before, after, or never), first gastroenteritis NOT at 2 yo (could be before, after, or never), mom with history of asthma → **Risk**

7.5 Stage 5: Feedback

The quality of the main predictive models was notable, even though the performance was not at a level to become usable as standalone predictive models in production. Therefore, we proceeded to evaluate the validity of findings. As the main predictive models did not have sufficient quality to become usable in production, we did not attempt to upgrade the models to a lifelong learning version. That was also not the goal of this experiment.

7.6 Limitations

The limitations in conducting this case study were the same as the ones mentioned for the first case study in section 6.6 as the system conditions to run experiments, source data, and selected variables were similar.

8 CASE STUDY III:

INFLAMMATORY BOWEL DISEASE &

SYSTEMIC AUTOIMMUNE RHEUMATIC DISEASES

The results from the first and second case studies were very similar. This could have happened because more than 97% of immune-mediated inflammatory disease (IMID) cases were people with asthma. However, this comparison still could not demonstrate whether those findings were applicable to all IMIDs or were relevant to only predicting patients with asthma. Therefore, we decided to conduct the same experiments on patients with IMIDs other than asthma. For the third case study, included both inflammatory bowel disease (IBD) and systemic autoimmune rheumatic disease (SARD) patients to investigate whether the findings resulting from the first two case studies were relevant to all IMIDs or were primarily driven by asthma.

In this chapter, we discuss our implementation of the designed methodology to predict Ontario newborns with a high risk of developing IBD or SARDs later in life (by 10 years of age) and generate new hypotheses of multifactorial association rules using a case-control study. The primary outcome is the incidence of IBD or SARDs. Therefore, only patients with IBD or SARDs were considered as cases. If the results were found to be similar to those from the first and second case studies, it would mean the findings from those two case studies were possibly relevant to all IMIDs. The goal was to build high-quality predictive models to identify significant factors, multifactorial rules, and a model that can predict patients at risk of IBD or

SARDs to compare with the results from the first and second case studies. The details of the process to conduct this case study by applying the stages of the reliable methodology, including preparation of the IBD and SARD datasets, characteristics of the datasets, list of variables, verification of potential risks in the data and study, approaches to modelling, performance of models, repetition of modelling, extracted findings, and evaluation of results, are presented in the following representative subsections. Note that we used health administrative data available at ICES to conduct this case study; therefore, no data collection occurred, as the health administrative data had already been gathered.

8.1 Stage 1: Data Selection

We had already gathered a list of risk, protective, and no effect factors associated with IBD and SARDs based on literature and domain experts' opinion; selected the required datasets available from ICES; and chosen attributes available in those datasets related to the collected factors in the first case study.

8.1.1 Factors

In section 6.1.1, we discussed the procedure and sources to collect the 128 factors in various categories. A full list of retrieved factors can be seen in Appendix I (IMIDs Early Life and Environmental Factors). Note that we included all collected factors in this case study, even if they were observed to be associated with other IMIDs, as it is possible these diseases have common root causes due to their similar characteristics and the dysregulations they cause in the body.

8.1.2 Datasets

We used the same selection of 88 datasets from different data libraries at ICES as described in section 6.1.2 for this case study. Even though some datasets contained information related to specific IMIDs, such as the ASTHMA dataset for asthmatic patients, we did not exclude them in this stage as they still could include variables useful for investigations on the results. Further information on the data selection procedures including the privacy assessment, ethics approvals, and cohort definitions are provided in the same section.

8.1.3 Attributes

As reported in section 6.1.3, 142 attributes from the selected datasets plus 288 attributes in the PM2.5 and NO2 datasets were selected. They eventually supported 67 factors from the collected list of factors. We assigned a type to each attribute to demonstrate the role it can play in preparing the final dataset from among these options: Exposure, Outcome, Linkage, Intermediary, and Auxiliary.

8.2 Stage 2: Preprocessing

Similar to the second case study, we could not use the IMM dataset as the prepared dataset in the first case study contained variables that were constructed based on each individual's outcome status. However, most preprocessing actions described in Appendix V (Preprocessing Details of the First Case Study) were still applicable as the same data were selected to be preprocessed, but for different outcomes. We first prepared the IBD dataset to analyze IBD patients, then we prepared the SARD dataset as the results from the IBD experiment were not promising. In the following sections, we summarize each step in

preprocessing both datasets and point out the differences compared to preprocessing actions taken in the first case study.

8.2.1 IBD Dataset

8.2.1.1 Technical Preprocessing

We named the dataset in which IBD was the target variable and the children with IBD formed its cases, the IBD dataset which initially contained 97 variables. As a result of the technical preprocessing actions, the **IBD dataset** was prepared containing **98 variables** (83 input, 1 target, 14 auxiliary variables) and **722,038 records** (split to 216,611 as the test set and 505,427 as the training set, which was balanced to 234 records using undersampling, 819 records using SMOTE, and 1,010,620 records using oversampling for modelling) after preprocessing. We have summarized the technical preprocessing actions to prepare the IBD dataset in Table 8.1. The ones different from the first case study are underlined.

The list of 83 predictors and one target variable used in these modelling experiments along with their descriptions is provided in Table 8.2.

8.2.1.2 Contextual Preprocessing

The potential systematic errors in conducting this case study were similar to those identified and described in the first case study, which are discussed in Appendix V. As an input variable reflected the prevalence of all IMIDs among mothers, the concerns in identifying IMID patients other than those with IBD were applicable in information biases but not in selection biases. We have listed the contextual preprocessing assessments that aimed to detect systematic errors in Table 8.3. The ones different from those in the first case study are underlined.

Table 8.1: Technical preprocessing actions in the third case study (IBD)

Step	*Title*	*Actions*
1	Dataset Creation	Base dataset: BORN-Niday After inclusion criteria: 813,719 records After exclusion criteria: 725,630 records
2	Feature Construction	Built 58 new features
3	Exploration and Planning	Init IBD dataset: 82 input, 1 target, 14 aux variables *{included M_IMMUNE_PRV in addition to M_IBD_PRV in inputs; and an aux variable showing the latest type of IBD – Crohn's vs UC}*
4	Data Partitioning	70/30 data split
5	Data Cleaning	Fixed values in BWEIGHT and INTBF variables
6	Missing Values	Replaced in FSRCE Removed 5 variables: BMI, HAD_INFLUENZA, MATWGTKG, RECEIVED_ANTIVIRAL, VACCINCATION_INFLUENZA Truncated 0.5% of records [After step 9:] Imputed 2.1% of cells *{this random selection of training records contained fewer missing cells}*
7	Categorical Variables	One-hot encoded: from 77 input variables to 591*
8	Feature Reduction	Removed 504 almost-unary variables Removed 4 highly correlated variables: B_SEX → SEX FSRCE.0 → M_IS_IMMIGRANT RIO2008_BIRTH → RIO2008_PREG ATOPY_IBD.NA → RHINITIS_IBD.NA
9	Data Normalization	Standardized all input values
10	Data Balancing	Undersampled: each class 117 records SMOTE: 351 records class-1 / 468 records class-0 Oversampled: each class 505,310 records *{there are much fewer cases; various experiments}*
11	Resampling Data	10-fold cross validation
12	Runtime Filtering	None

** The number of one-hot encoded variables were slightly more than those produced in preprocessing the ASTH dataset. CYTO.8, RETRO.5, and PSEUDO.9 were the variables that were not produced in ASTH. The reason is that there are patients who developed these comorbidities for the first time at the mentioned ages, but most likely were diagnosed with asthma before developing these comorbidities. Therefore, their values for these comorbidities were "not applicable" in the ASTH dataset.*

Table 8.2: List of predictors and target variable in IBD models

Name	Label	Purpose
sex	*Sex*	Factor: sex
gest	*Gestational age at birth (weeks)*	Factor: late premature birth
deltype	*Delivery type*	Factor: c-section
intbf	*Intention to breastfeed*	Factor: breastfeeding, cow's milk formula (doesn't show mixed feeding, so if the value is TRUE the baby may still have taken formula)
labtype.(spo,ind,no)	*Labour type*	Factor: complicated delivery (receiving induction)
mathp	*Maternal health problem(s)*	Effects of maternal problems
nbres	*Newborn Resuscitation*	Effects of reviving newborn
obcomp	*Obstetrical complications*	Factor: complicated delivery
parity	*Parity*	Factor: firstborn child, household crowding, personal towel
smoking.(no,yes)	*Smoking (during preg)*	Factor: maternal smoking, passive smoking exposure
b_bmonth.(1-12)	*Baby's birth month*	Factor: fungal spore season, grass pollen exposure (Apr to Jul), sunlight exposure
m_bmonth.(1-12)	*Mother's birth month*	Effects of mother's birth month
assisted.(no,vac)	*Forceps / vacuum*	Factor: complicated delivery (assisted tools)
augment	*Augmentation*	Factor: complicated delivery (labour augmentation)
bweight	*Birth weight (grams)*	Factor: low birth weight
B_IBD_INC	*Child's incidence of IBD*	Outcome
M_IMMUNE_PRV	*Mother's prevalence of any immune disease*	Factors related to family history
incquint_birth.(1-5)	*Nearest Census Based Neighbourhood Income Quintile at birth*	Factor: high-income lifestyle
incquint_PREG.(1-5)	*Nearest Census Based Neighbourhood Income Quintile at conception*	Factor: high-income lifestyle

rio2008_PREG	*2008 Rurality Index for Ontario at conception*	Factor: living in farm, urbanization
M_age	*Mother's age at birth*	Factor: maternal age
fsrce.(asia/pac)	*Mother's source area*	Mother/family's origin
M_IS_IMMIGRANT	*Mother is immigrant*	Factor: 2nd-gen immigrant
pm25_birth	*Air pollution level of PM2.5 in region of residence at birth*	Factor: air pollution, traffic-related particles, fine particulates
pm25_preg	*Air pollution level of PM2.5 in region of residence during pregnancy*	Factor: air pollution, traffic-related particles, fine particulates
no2_birth	*Air pollution level of NO2 in region of residence at birth*	Factor: air pollution, traffic-related particles
no2_preg	*Air pollution level of NO2 in region of residence during pregnancy*	Factor: air pollution, traffic-related particles
Gen_Child_Inf_IBD .(some ages)	*Child diagnosed with general childhood infections before incidence of IBD*	Factor: general childhood infection
Antibiotics_IBD .(some ages)	*Child took antibiotics before incidence of IBD*	Factor: use of antibiotics
Gastro_inf_IBD .(some ages)	*Child diagnosed with gastroenteritis before incidence of IBD*	Factor: gastroenteritis
Respir_inf_IBD .(some ages)	*Child diagnosed with respiratory infection before incidence of IBD*	Factor: respiratory infection
Rhinitis_IBD.na	*Child diagnosed with allergic rhinitis before incidence of IBD*	Factor: rhinitis
Strep_pyo_IBD.na	*Child diagnosed with Streptococcus pyogenes before incidence of IBD*	Factor: streptococcus pyogenes bacteria

Table 8.3: Contextual preprocessing assessments in the third case study (IBD)

Step	*Title*	*Assessments*
1	Selected Source Databases	*Selection bias:* • Low risk of ascertainment bias as the sensitivity and specificity in source databases were not 100%, but still high.
2	Criteria of Study Population	*Selection bias:* • Low risk of inclusion bias as babies born out of hospitals were not included (<3% of births). • Low risk of exclusion bias as children who moved out-of-province were excluded. • No matching bias.
3	Diagnostic Algorithms and Data Completeness	*Selection bias:* • No risk of diagnostic bias as diabetic patients were not part of the cases so that lack of type of diabetes in source data could cause an issue. • No risk of prevalence-incidence bias as the incident information of cases were available. *Information bias:* • Medium risk of misclassification bias due to inaccuracy in diagnostic algorithms and underlying data to construct comorbidity variables. • Low risk of diagnostic bias as the type of diabetes was not identified in the source data to construct mother's history variables. • Low risk of prevalence-incidence bias as prevalent information of mothers was used to construct outcome history variables.
4	Use of Diagnostic and Procedural Codes	*Information bias:* • Low risk of misclassification bias due to constructing features using various diagnostic and procedural coding systems.
5	Time-Varying Variables	*Selection bias:* • Medium risk of diagnostic bias in identifying IBD patients due to short follow up period. • Low risk of survivor treatment bias in selecting controls. *Information bias:* • Low risk of immortal time bias in constructing comorbidity variables in controls. • Low risk of misclassification bias in constructing location-based variables during pregnancy time.

6	Lack of Contributing Variables	*Confounding:* • High risk of unmeasured confounding since clinical factors were not supported and dads' information was not accessible. • Low risk of measured confounding due to the elimination of mostly almost-unary variables containing limited information.
7	Handled Missing Values	*Selection bias:* • Trivial risk of exclusion bias as 0.5% of individuals were excluded. *Confounding:* • Medium risk of measured confounding due to elimination of few variables representing various factors due to high missingness.
8	Partitioned, Balanced, and Resampled Data	*Selection bias:* • Trivial risk of non-random sampling bias since random selection was used in all required sampling attempts. • No risk of matching bias as no matching method was used to do sampling.
9	Model Results	*Selection bias:* • Trivial risk of publication bias as all findings from models with reliable performance were reported.

8.2.2 SARD Dataset

8.2.2.1 Technical Preprocessing

We named the dataset in which SARDs was the target variable and the children with any SARD formed its cases, the SARD dataset which initially contained 96 variables. As a result of the technical preprocessing actions, the **SARD dataset** was prepared containing **98 variables** (84 input, 1 target, 13 auxiliary variables) and **722,038 records** (split to 216,611 as test set and 505,427 as training set, which was balanced to 3,422 records — 2 times 1,711 — for modelling) after preprocessing. We have summarized the technical preprocessing actions to prepare the SARD dataset in Table 8.4. The ones different from those for the first case study are underlined.

Table 8.4: Technical preprocessing actions in the third case study (SARD)

Step	*Title*	*Actions*
1	Dataset Creation	Base dataset: BORN-Niday After inclusion criteria: 813,719 records After exclusion criteria: 725,630 records
2	Feature Construction	Built 58 new features
3	Exploration and Planning	Init SARD dataset: 82 input, 1 target, 13 aux variables *{included M_IMMUNE_PRV in addition to M_SARD_JIA_PRV in inputs}*
4	Data Partitioning	70/30 data split
5	Data Cleaning	Fixed values in BWEIGHT and INTBF variables
6	Missing Values	Replaced in FSRCE Removed 5 variables: BMI, HAD_INFLUENZA, MATWGTKG, RECEIVED_ANTIVIRAL, VACCINCATION_INFLUENZA Truncated 0.5% of records [After step 9:] Imputed 2.1% of cells *{this random selection of training records contained fewer missing cells}*
7	Categorical Variables	One-hot encoded: from 77 input variables to 591*
8	Feature Reduction	Removed 503 almost-unary variables Removed 4 highly correlated variables: B_SEX → SEX M_IS_IMMIGRANT → FSRCE.0 RIO2008_BIRTH → RIO2008_PREG ATOPY_SARD.NA → RHINITIS_SARD.NA
9	Data Normalization	Standardized all input values
10	Data Balancing	Undersampled: each class 1,711 records *{there are much fewer cases}*
11	Resampling Data	10-fold cross validation
12	Runtime Filtering	None

** Same description mentioned under Table 8.1*

The list of 84 predictors and one target variable used in these modelling experiments along with their descriptions is provided in Table 8.5.

Table 8.5: List of predictors and target variable in SARD models

Name	Label	Purpose
sex	*Sex*	Factor: sex
gest	*Gestational age at birth (weeks)*	Factor: late premature birth
deltype	*Delivery type*	Factor: c-section
intbf	*Intention to breastfeed*	Factor: breastfeeding, cow's milk formula (doesn't show mixed feeding, so if the value is TRUE the baby may still have taken formula)
labtype.(spo,ind,no)	*Labour type*	Factor: complicated delivery (receiving induction)
mathp	*Maternal health problem(s)*	Effects of maternal problems
nbres	*Newborn Resuscitation*	Effects of reviving newborn
obcomp	*Obstetrical complications*	Factor: complicated delivery
parity	*Parity*	Factor: firstborn child, household crowding, personal towel
smoking.(no,yes)	*Smoking (during preg)*	Factor: maternal smoking, passive smoking exposure
b_bmonth.(1-12)	*Baby's birth month*	Factor: fungal spore season, grass pollen exposure (Apr to Jul), sunlight exposure
m_bmonth.(1-12)	*Mother's birth month*	Effects of mother's birth month
assisted.(no,vac)	*Forceps / vacuum*	Factor: complicated delivery (assisted tools)
augment	*Augmentation*	Factor: complicated delivery (labour augmentation)
bweight	*Birth weight (grams)*	Factor: low birth weight
B_SARD_JIA_INC	*Child's incidence of SARD + JIA*	Outcome
M_SARD_JIA_PRV	*Mother's prevalence of SARD + JIA*	Factors related to family history
M_IMMUNE_PRV	*Mother's prevalence of any immune disease*	Factors related to family history
incquint_birth.(1-5)	*Nearest Census Based Neighbourhood Income Quintile at birth*	Factor: high-income lifestyle

incquint_PREG.(1-5)	*Nearest Census Based Neighbourhood Income Quintile at conception*	Factor: high-income lifestyle
rio2008_PREG	*2008 Rurality Index for Ontario at conception*	Factor: living in farm, urbanization
M_age	*Mother's age at birth*	Factor: maternal age
fsrce.(no,asia/pac)	*Mother's source area*	Mother/family's origin
pm25_birth	*Air pollution level of PM2.5 in region of residence at birth*	Factor: air pollution, traffic-related particles, fine particulates
pm25_preg	*Air pollution level of PM2.5 in region of residence during pregnancy*	Factor: air pollution, traffic-related particles, fine particulates
no2_birth	*Air pollution level of NO2 in region of residence at birth*	Factor: air pollution, traffic-related particles
no2_preg	*Air pollution level of NO2 in region of residence during pregnancy*	Factor: air pollution, traffic-related particles
Gen_Child_Inf_SARD_JIA .(some ages)	*Child diagnosed with general childhood infections before incidence of SARD + JIA*	Factor: general childhood infection
Antibiotics_SARD_JIA .(some ages)	*Child took antibiotics before incidence of SARD + JIA*	Factor: use of antibiotics
Gastro_inf_SARD_JIA .(some ages)	*Child diagnosed with gastroenteritis before incidence of SARD + JIA*	Factor: gastroenteritis
Respir_inf_SARD_JIA .(some ages)	*Child diagnosed with respiratory infection before incidence of SARD + JIA*	Factor: respiratory infection
Rhinitis_SARD_JIA.na	*Child diagnosed with allergic rhinitis before incidence of SARD + JIA*	Factor: rhinitis
Strep_pyo_SARD_JIA.na	*Child diagnosed with Streptococcus pyogenes before incidence of SARD + JIA*	Factor: streptococcus pyogenes bacteria

8.2.2.2 Contextual Preprocessing

The potential systematic errors in conducting this case study were similar to those identified and described in the first case study, which are discussed in Appendix V (Preprocessing Details of the First Case Study). As an input variable reflected the prevalence of all IMIDs among mothers, the concerns in identifying IMID patients other than those with SARD were applicable in information biases but not in selection biases. In Table 8.6, we have listed the

contextual preprocessing assessments that aimed to detect systematic errors. The ones different from those in the first case study are underlined.

Table 8.6: Contextual preprocessing assessments in the third case study (SARD)

Step	*Title*	*Assessments*
1	Selected Source Databases	*Selection bias:* • Low risk of ascertainment bias as the sensitivity and specificity in source databases were not 100%, but still high.
2	Criteria of Study Population	*Selection bias:* • Low risk of inclusion bias as babies born out of hospitals were not included (<3% of births). • Low risk of exclusion bias as children who moved out-of-province were excluded. • No matching bias.
3	Diagnostic Algorithms and Data Completeness	*Selection bias:* • No risk of diagnostic bias as diabetic patients were not part of the cases so that lack of type of diabetes in source data could cause an issue. • No risk of prevalence-incidence bias as the incident information of cases were available. *Information bias:* • Medium risk of misclassification bias due to inaccuracy in diagnostic algorithms and underlying data to construct comorbidity variables. • Low risk of diagnostic bias as the type of diabetes was not identified in the source data to construct mother's history variables. • Low risk of prevalence-incidence bias as prevalent information of mothers was used to construct outcome history variables.
4	Use of Diagnostic and Procedural Codes	*Information bias:* • Low risk of misclassification bias due to constructing features using various diagnostic and procedural coding systems.
5	Time-Varying Variables	*Selection bias:* • No risk of diagnostic bias in identifying IBD patients due to short follow up period as they are not part of cases. • Low risk of survivor treatment bias in selecting controls.

		Information bias: • Low risk of immortal time bias in constructing comorbidity variables in controls. • Low risk of misclassification bias in constructing location-based variables during pregnancy time.
6	Lack of Contributing Variables	*Confounding:* • High risk of unmeasured confounding since clinical factors were not supported and dads' information was not accessible. • Low risk of measured confounding due to the elimination of mostly almost-unary variables containing limited information.
7	Handled Missing Values	*Selection bias:* • Trivial risk of exclusion bias as 0.5% of individuals were excluded. *Confounding:* • Medium risk of measured confounding due to elimination of few variables representing various factors due to high missingness.
8	Partitioned, Balanced, and Resampled Data	*Selection bias:* • Trivial risk of non-random sampling bias since random selection was used in all required sampling attempts. • No risk of matching bias as no matching method was used to do sampling.
9	Model Results	*Selection bias:* • Trivial risk of publication bias as all findings from models with reliable performance were reported.

8.3 Stage 3: Modelling

As our target variable was categorical, we built models using classification algorithms first to predict whether a new patient is at high risk of developing IBD and then similarly for SARDs, and to identify the most important variables and multifactorial rules that cause a patient to be predicted as high risk, which generates new hypotheses of factors associated to these outcomes.

Similar to the second case study, we used the same three classifiers from the first case study in order to compare the results. The selected classifiers were the C5 decision tree, single hidden layer neural network, and logistic regression. In addition, we used AUC (area under the ROC curve) as the main metric to assess the quality of the predictive models throughout the entire process and considered sensitivity as well due to the same reasons explained in the first case study.

We built predictive models using the selected classifiers on the preprocessed (including data balancing) IBD training set and then on the preprocessed SARD training set. Even though most models on the IBD dataset performed very poorly, the C5 Tree on the undersampled training set provided slightly better results. Therefore, we were able extract multifactorial rules from the C5 decision tree. Nevertheless, we decided to focus on another IMID, excluding asthma, with a number of cases higher than that of IBD, for which we selected SARDs. While the C5 Tree performed better in predicting SARD cases, both the NNet and GLM_log performed better in predicting controls and gained a higher overall quality. We also extracted multifactorial rules from the C5 decision tree on the SARD dataset.

The attributes were ranked as each predictive model was trained. We only had access to the importance score of the top 20 variables in each model due to the limitation in the selected R package, so we calculated the adjusted score of the variables to identify those that were important across all SARD predictive models.

8.4 Stage 4: Evaluation

Below, we report the results of modelling experiments and discuss the quality of the models. Then, we present the findings from models with satisfactory quality and discuss their validity.

8.4.1 Models

In the predictive models, we expected the AUC and sensitivity on unseen data to not drop significantly compared to the results from model training. If they did, then the model was overfit and not usable. Also, we expected the model's sensitivity to be acceptable so that we could rely on it. As our study is in the health context, we prefer that more controls are predicted as cases rather than more cases being predicted as controls. The latter is dangerous, and we wanted to avoid it as much as possible.

We built predictive models using the selected classifiers on the IBD dataset. First, the models were fitted using the undersampled training set, then with the training set that was balanced using SMOTE, and finally, the oversampled training set. The details and results of these models are reported in Table 8.7.

Table 8.7: Details and results of predictive modelling on the IBD dataset

Imbalance	*class-1 [0.02%]*	117	Predictors	Predict IBD using C5 Tree		Predict IBD using default NNet		Predict IBD using binomial GLM	
	class-0	505,310	83	*Train*	*Test*	*Train*	*Test*	*Train*	*Test*
Undersampling	*class-1*	117	AUC	0.6875	0.6586	0.6120	0.5487	0.5706	0.5815
	class-0	117	Sens	0.7523	0.7400	0.5038	0.5800	0.4939	0.5600
	Samples	234	Spec	0.5811	0.5773	0.5326	0.5174	0.4780	0.6029
SMOTE	*class-1*	351	AUC	0.8983	0.5345	0.8907	0.4752	0.7660	0.6062
	class-0	468	Sens	0.7671	0.1400	0.9263	0.1800	0.6750	0.4600
	Samples	819	Spec	0.8846	0.9289	0.7328	0.7704	0.7390	0.7524
Oversampling	*class-1*	505,310	AUC	1	0.5	0.9704	0.5051	0.8100	0.6071
	class-0	505,310	Sens	1	0	0.9648	0.0800	0.7830	0.5000
	Samples	1,010,620	Spec	0.99	1	0.9224	0.9302	0.7130	0.7141

We interpreted the results as follows:

- The AUC of the C5 Tree model on the undersampled data was around 68% — similar to the AUC of models in the first and second case studies — with an acceptable sensitivity of 75%. However, more than 40% of the full test set were erroneously predicted as cases (FP). This is a massive error and cannot help much in properly identifying cases. The confusion matrix on the test set was TP:37; FN:13; TN: 125,018; FP: 91,543.
- The results of both the neural network and logistic regression models on the undersampled data were even worse than the C5 Tree. These models were acting more like a random coin flip.
- Both the C5 Tree and the neural network on the SMOTE data were totally overfit as they performed very well on the training set but had terrible prediction on unseen data.
- The logistic regression on the SMOTE data was also overfit; however, it was not as bad as the other two SMOTE models, while still not predicting the cases very well.
- Again, both the C5 Tree and the neural network on the oversampling data were totally overfit and useless. The accuracy on both the training and test sets was significantly misleading (the accuracy of the C5 Tree was 0.99), which again demonstrated why AUC is a more reliable metric to evaluate the quality performance of predictive models.
- Similar to the SMOTE result, the logistic regression on the oversampled data performed better, but it was still overfit, performing with complete randomness in predicting cases.

As most IBD models performed poorly and one model in each round was relatively outstanding, we did not run the test of significance between the models. However, we did identify the variables highly contributing to the models that performed better compared to the other models using the same balancing method. Table 8.10 shows the list of these variables in predicting children with high risk of developing IBD. The diagnosis of gastroenteritis was the common highly important variable. The label and description of these variable names are mentioned in Table 8.2 and Appendix IV.

Table 8.8: Top contributing variables in IBD models with better performance in the third case study

C5 Tree on Undersampled Data	Binomial GLM on SMOTE Data	Binomial GLM on Oversampled Data
b_bmonth.(3,6,7,9,12)	*gastro_inf.na*	*gastro_inf.na*
gen_child_inf.1	strep_pyo.na	strep_pyo.na
gastro_inf.na		
respir_inf.na		
intbf		

To continue, we built predictive models using the selected classifiers on the SARD dataset. The details and results of these models are reported in Table 8.9.

Table 8.9: Details and results of predictive modelling on the SARD dataset

Goal		Predict SARDs using C5 Tree		Predict SARDs using default NNet		Predict SARDs using binomial GLM	
Run	System	*type*	ICES Intranet: R job	*type*	ICES Intranet: R job	*type*	ICES Intranet: R job
	Duration	*hours*	92.2	*hours*	92.3	*hours*	92.1
	Costly	*impute training*	65h 30m	*impute training*	65h 30m	*impute training*	65h 30m
		prep test	26h 26m	*prep test*	26h 26m	*prep test*	26h 26m
		fitting	3m	*fitting*	9m	*fitting*	10s
Training Set	Imbalance	*class-1 [org:0.34%]*	1,711	*class-1 [org:0.34%]*	1,711	*class-1 [org:0.34%]*	1,711
		class-0	503,716	*class-0*	503,716	*class-0*	503,716
	Balance	Undersampled		Undersampled		Undersampled	
	Target	*class-1*	1,711	*class-1*	1,711	*class-1*	1,711
		class-0	1,711	*class-0*	1,711	*class-0*	1,711
Test Set		Preprocessed		Preprocessed		Preprocessed	
		Kept imbalanced		Kept imbalanced		Kept imbalanced	
Modelling	Classifier		C5		Neural Network		GLM
	Resampling		CV-10		CV-10		CV-10
	Metric		ROC		ROC		ROC
	Tuning Parameters	*model*	tree	*size*	1, 3, 5, 10		
		winnow	FALSE, TRUE	*decay*	0, 1e-4, 0.1, 0.5		
		trials	1, 10, 20				
	Samples		3,422		3,422		3,422
	Predictors		84		84		84
	Classes		2		2		2
Best Model Tune	Parameters	*model*	tree	*size*	1		
		winnow	FALSE	*decay*	0.5		
		trials	10				
	Specific Results	*# boosting iterations*	10	*# iterations*	100	*degrees of freedom*	3421
		avg tree size	5.7	*network*	84-1-1	*AIC*	3941
		# leaves in first tree (no boosting)	5	*weights*	87	*# fisher iteration*	4

	AUC		**0.7328**		**0.7597**		**0.7615**
	Sens		**0.6154**		**0.6271**		**0.6616**
	Spec		0.7364		0.7680		0.7241
Performance on Unseen	AUC (fitting)		0.6709		0.6519		0.6929
	Confusion Matrix	*1 -> 1 [exp: 0.34%]*	517	*1 -> 1 [exp: 0.34%]*	459	*1 -> 1 [exp: 0.34%]*	487
		1 -> 0	216	*1 -> 0*	274	*1 -> 0*	246
		0 -> 1	82,160	*0 -> 1*	55,493	*0 -> 1*	62,955
		0 -> 0 [exp: 99.7%]	133,718	*0 -> 0 [exp: 99.7%]*	160,385	*0 -> 0 [exp: 99.7%]*	152,923
	Acc		0.6197		0.7425		0.7082
	AUC		**0.6624**		**0.6846**		**0.6864**
	Sens		**0.7053**		**0.6262**		**0.6644**
	Spec		0.6194		0.7429		0.7084

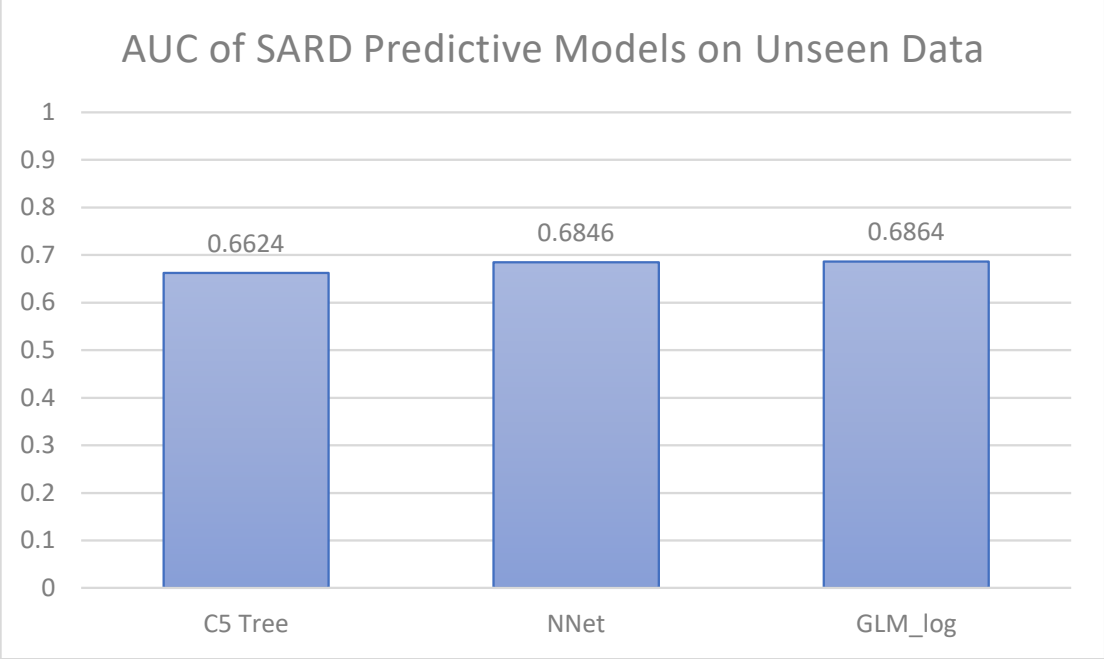

Figure 8.1: Performance of SARD predictive models on unseen data

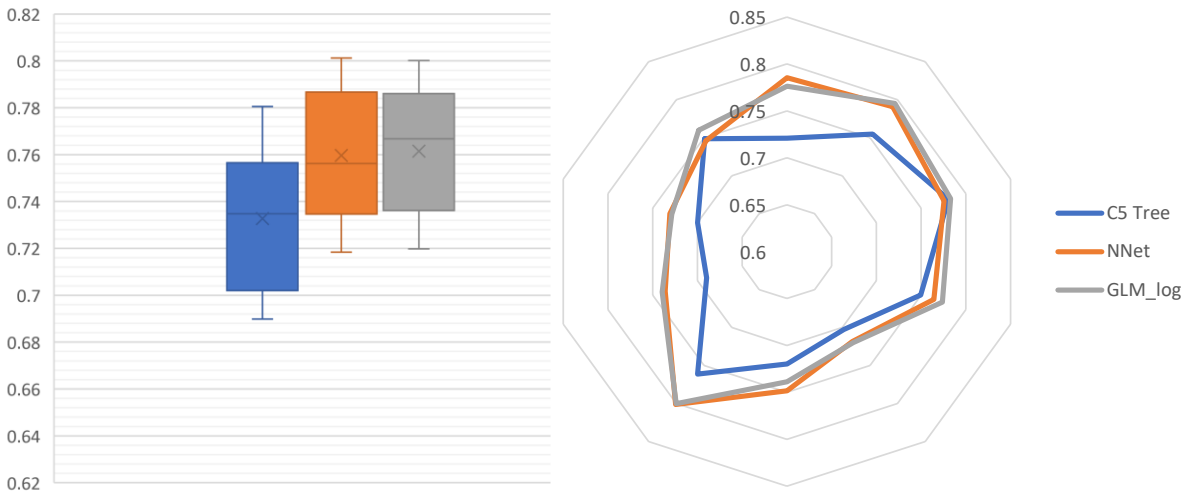

Figure 8.2: Performance of SARD predictive models during 10-fold cross validation

Observing many interesting and fluctuating results in these models, we interpreted them as follows:

- The C5 Tree predicted cases better and underperformed in predicting controls compared to the other two models. Therefore, its accuracy is lower than its AUC as the controls form the larger class. Interestingly and in contrast, the estimated sensitivity on the training set was lower than the specificity. This could have happened because the controls were significantly undersampled causing the model to miss some patterns among the controls which then showed up in the unseen data.
- In contrast to the C5 Tree, both the neural network and logistic regression predicted controls with higher quality than the cases. Therefore, their accuracy was higher than their AUC.
- The models were not overfit and performed similarly on unseen data as on the training set. The AUCs are shown in Figure 8.1. This was in favor of the models' usability.
- The AUC and sensitivity of 62 to 70% are notable as they represent only weak to moderate prediction. In other words, we could extract some interesting information,

but the models could not be used for real-world prediction solutions as they did not have acceptable performance.

We ran the test of significance between the three models to identify which performed better. The difference between the neural network and logistic regression was not statistically significant ($d = 0.002 \pm 2.26 \times 0.002$; CV-10, CL=0.95) since the confidence interval did not span the value zero at a 95% confidence level in estimating the performance using a 10-fold cross-validation approach. However, the C5 Tree had a significant difference from both the neural networks ($d = 0.027 \pm 2.26 \times 0.007$; CV-10, CL=0.95) and the logistic regression ($d = 0.029 \pm 2.26 \times 0.005$; CV-10, CL=0.95), with a slightly lower performance. The differences in performance can also be seen in Figure 8.2.

The adjusted importance scores of the variables were calculated to identify the important variables across all three models. Table 8.10 shows the list of the top 20 important variables in predicting children with high risk of developing SARDs. Note that the 11 variables in green were observed among the top 20 important variables in all models. The label and description of these variable names are mentioned in Table 8.5 and Appendix IV, and we also elaborate them when evaluating the findings in section 8.4.2.3.

8.4.2 Findings

The findings from models with satisfactory quality are presented and discussed in this section. We also compare the findings of the first and third case studies. It seems the results of predicting SARDs are slightly different from the first case study. In other words, the difference between the IMM and ASTH results was much lower than what we observe between the IMM and SARD results. This finding confirms that the results of the first case study were dominated

by asthmatic patients who formed more than 97% of the IMID cohort (Figure 6.2). The comparison of each finding is discussed in the following sections.

Table 8.10: Top 20 important variables in the third case study

Variable	*Sum*	*App*	*Avg*	*Weighted Avg*
antibiotics.na	257.93	3	85.98	148.92
m_sard_jia_prv	247.73	3	82.58	143.03
strep_pyo.na	246.93	3	82.31	142.57
gastro_inf.na	229.93	3	76.64	132.75
gastro_inf.3-12m	223.85	3	74.62	129.24
gen_child_inf.nb	222.14	3	74.05	128.25
respir_inf.na	202.47	3	67.49	116.9
gen_child_inf.na	200.28	3	66.76	115.63
rhinitis.na	146.12	2	73.06	103.32
respir_inf.3-12m	144.22	2	72.11	101.98
antibiotics.2	100	1	100	100
antibiotics.4	100	1	100	100
gen_child_inf.1-3m	100	1	100	100
no2_preg	165.4	3	55.13	95.49
gen_child_inf.3-12m	90.32	1	90.32	90.32
sex	145.03	3	48.34	83.73
respir_inf.2	83.23	1	83.23	83.23
respir_inf.1	116.97	2	58.49	82.71
gen_child_inf.1	76.94	1	76.94	76.94
antibiotics.3-12m	106.41	2	53.21	75.24
respir_inf.1-3m	127.08	3	42.36	73.37

8.4.2.1 Considerations

In addition to the quality of the models, which we evaluated in the previous section, the level of systematic error plays a significant role in having reliable findings. Table 8.3 and Table 8.6 summarized the level of risks in this regard. We tolerated the trivial risks, but there were still

some higher risks of bias and confounding which we needed to take into consideration. To summarize, we need to consider the following risks and limitations along with results and findings of this case study:

- **Confounding:** The results come with high risk of unmeasured confounding and medium risk of measured confounding as we could not include clinical factors and fathers' information. We also had to eliminate a few factors due to high missing data and many factors with almost-unary values.
- **Information bias:** The results come with high risk of misclassification bias and low risk of diagnostic bias, prevalence-incidence bias, and immortal time bias as there were inaccuracy in mothers' postal codes at conception time, where were used to construct location-based variables; significant missing values in air quality raw data, which affects calculating the average value; and inaccuracy in diagnostic algorithms and underlying data to construct comorbidity variables. Moreover, we did not have the diabetes type in the source data for mothers' histories and also had to use prevalence information of mothers to construct outcome history variables, various diagnostic and procedural coding systems to construct comorbidity variables, and different index dates to construct comorbidity variables in controls.
- **Selection bias:** The results come with low risk of inclusion bias, exclusion bias, and survivor treatment bias as we could not include babies born out of hospitals, had to exclude children who had moved out of the province, and did not have a significantly long follow-up window in selecting controls. In addition, the results of the IBD models come with medium risk of diagnostics risk as we had a short follow-up period in identifying IBD patients.

8.4.2.2 Prediction Model

One of the findings of this case study was the predictive model itself as it could be used in production for real-world prediction solutions such as an alerting system which detects high-risk patients. The IBD models were either overfit or had very poor performance, except for the C5 Tree on undersampled data. However, the SARD models we built were not overfit and produced proper predictions. Their AUCs were 68%, which is considered to be weakly or moderately predictive. Therefore, the models could be used to extract new information, but they could not be used in production.

Even though the overall performance of predictive models in both case studies are very similar, the SARD models did not predict cases and controls with equal quality. Furthermore, in the first case study, the logistic regression model performed slightly worse than the other two models, while the C5 Tree model performed slightly worse than the other two models in the third case study. Overall, it can be observed that the model behaviours have slightly changed and there are differences in predicting cases vs controls between the IMM and SARD models, compared to the IMM and ASTH models that showed quite similar behaviors.

8.4.2.3 Significant Factors

As most IBD models did not have satisfactory quality, we only identified the highly contributing variables in the models with slightly better performance, as shown in Table 8.8. The only variable that was common among the selected models was “gastroenteritis”.

Another finding of this case study was the important variables that were among the top 20 of all three SARD models. These variables had the highest association in predicting SARD children:

- General childhood infections, especially with first time at newborn age
- Streptococcus pyogenes
- Use of antibiotics
- Respiratory infection, especially with first episode at 1–3 months of age
- Mother's prevalence of SARDs
- Gastroenteritis, especially with first episode at 3–12 months of age
- Air pollution level of NO2 in region of residence during pregnancy
- Baby's sex

These factors highly overlap with those of the first case study. Therefore, it seems that mostly the same factors have significant impact in identifying any IMID patient. However, in SARD models, having general childhood infection (no matter the age) and the air quality index also play an important role. However, we have reported there is high risk of misclassification bias in the variables of air quality index at residence during pregnancy that needs to be considered.

8.4.2.4 Multifactorial Rules

We extracted the paths on the decision tree that had a confidence score higher than or close to the models' performances. All the extracted rules with their related details are listed in Appendix VIII (for IBD) and Appendix IX (for SARDs). Further information on how to interpret these rules are also provided in section 6.4.2.4.

When we compare the extracted rules with higher confidence score and notable number of applicable individuals to the ones discovered in the first case study, we do not see the high similarity that was observed between first and second case studies. In the IBD and SARD decision trees, "gastroenteritis" and "mother's prevalence of SARDs" plays the main role in

all strong rules, respectively, while the “use of antibiotics” was the primary factor in the IMM and ASTH decision trees.

8.5 Stage 5: Feedback

The IBD models on the undersampled data did not have appropriate quality, except for the decision tree. We thought perhaps the small data size (234 training samples) was not allowing the models to properly find patterns. Therefore, we took this feedback and returned to the preprocessing stage to balance the IBD training set using the SMOTE technique. However, the new models performed worse, while the data size was still small. Finally, we again returned to the preprocessing stage and oversampled the IBD training set to balance the data. The quality of the new models was not also satisfactory. Therefore, evaluating their findings was not necessary and we started a new series of experiment to analyze patients with SARD as a non-asthma IMID.

The quality of the IBD decision tree model on undersampled data and all SARD models were satisfactory, even though the performance was not at a level to become usable as standalone predictive models in production. Therefore, we proceeded to evaluate the validity of findings. As these models did not have sufficient quality to become usable in production, we did not attempt to upgrade the models to a lifelong learning version. That was also not the goal of this experiment.

8.6 Limitations

The limitations in conducting this case study were the same as the ones mentioned for the first case study in section 6.6 as the system conditions to run experiments, source data, and selected variables were similar.

9 CONCLUSIONS

In this chapter, we summarize the objectives, achievements, and findings of this thesis work. In addition, we review the contributions and limitations of this study, and, finally, discuss potential future works.

9.1 Summary

Data mining includes non-trivial techniques that provide strong analytical abilities to discover meaningful and actionable hidden patterns in large amounts of data. Although data mining is widely applied in the health care sector, there is often doubt regarding the validity of analyses, which may be mistrusted by clinicians, epidemiologists, or public health specialists. As a matter of course while operating electronic health systems, large amounts of routinely collected health data are amassed, including health administrative data generated through administration of the health system. These data are not collected for research purposes; however, they are great sources of information for research as a secondary use of the data. The main objective of this research study was to design a solution to reliably analyze health administrative data using data mining techniques. It was an attempt to advance the dialogue between health experts and computer scientists and form common grounds and new collaborations to enhance the health analytical capabilities.

There are contextual considerations that are included in every epidemiological study in order to provide valid and reliable results. These are normally not taken into account in studies that apply data mining to health data. We defined two research questions to identify the missing

parts in order to provide trustworthy results and to design a reliable data mining methodology on health administrative data.

As we were motivated to shed light on the investigations around immune-mediated inflammatory diseases (IMIDs) when we commenced this research study, we defined a second objective to implement the designed reliable methodology on real-world data to run case studies on these diseases and evaluate the methodology. The root causes of these disorders are unknown. The rising rates of pediatric immune-mediated conditions in Canada and worldwide is alarming, and finding predisposing factors for these conditions is of high priority for Canadian pediatric health care providers. We focused on pediatric patients as our target population as they are exposed to fewer environmental factors. We defined a new set of research questions to address the second objective, which included implementing the designed reliable methodology on health administrative data to generate new hypotheses on the association of factors with IMIDs, and identifying the significant early life and environmental factors that predict risk of IMIDs.

9.2 Answers to the Research Questions

In this section, we summarize the answers to the research question defined for this study.

RQ1) What are the missing parts in a typical data mining study that are accounted for in many epidemiological studies in order to provide trustworthy results?

The missing parts to provide reliable results when applying data mining technique to health administrative data were encapsulated into two main components which we entitled "Contextual Preprocessing" and "Findings Evaluation." We understood that, in addition to making the data suitable to build a data mining model by running a series of technical

preprocessing actions on it to comply with the algorithm's requirements, what information is included in the data is remarkably important. This included investigating both the quality of data collection and the individuals included in the data.

RQ2) Can we design a reliable data mining methodology to produce valid findings when analyzing health administrative data while addressing the related epidemiological considerations?

We adopted the attempt to assess the risk of various sources of bias and embedded it in our designed data mining methodology. First, we verified which sources of bias are generally applicable to analyzing health administrative data and specifically relevant when data mining methods are applied. Then, we designed a series of steps to identify and mitigate the risk of each source of bias after the technical preprocessing is done. These steps were bundled in an organized fashion called the Contextual Preprocessing Guideline. Furthermore, in addition to building models with high-quality results, the findings extracted from the models should be validated by a domain expert. The findings are either supported by the literature or newly presented. If the former, it is a sign the results are valid; if the latter, the domain expert can help interpret whether there could be a clinical explanation or if there are contradictions with the domain knowledge. This way, the findings are validated and enhanced to provide reliable results.

The reliable data mining methodology on health administrative data was designed in this study to address impartiality, validity, and sustainability concerns in five stages: Data Selection, Preprocessing, Modelling, Evaluation, and Feedback. The Data Selection stage covered the factor selection (from the literature and the domain expert), dataset selection, and attribute selection procedures. The Preprocessing stage was formed of two guidelines — technical

preprocessing and contextual preprocessing — that were designed in a standard format specifically to apply data mining to health administrative data. The Modelling stage took care of building descriptive and predictive models. The Evaluation stage also consisted of two parts: the Models Evaluation, in which the results are presented and the quality of models is assessed, and the Findings Evaluation where the findings are presented and their validity is discussed. Finally, depending on the evaluations done on the model results and findings, the Feedback stage would lead to the Preprocessing or Modelling stages, or even upgrading the models to a lifelong learning version if all results are acceptable and the models are stable.

RQ3) How can this methodology be applied to health administrative data to generate new hypotheses on the association between IMIDs and early life and environmental factors?

Three case studies with different goals were defined to analyze children born in Ontario who developed IMIDs. The first case study focused on all IMIDs as the outcome because it was believed these diseases could have causes in common. The second case study considered asthmatic patients as cases because they formed the majority of cases of IMIDs in the cohort, and we attempted to verify whether they dominated the results in the first case study. As the results in the first two case studies were quite similar, the third case study focused on IBD and SARD as non-asthma outcomes to investigate whether the results change. The results of the case studies were interesting as some were supported by the literature and others raised questions about the reason for particular rules in the data.

The risk of applicable biases was also reported along with the results. The risk of unmeasured confounding due to inaccessible clinical data and measured confounding due to removing

factors during preprocessing were marked as high and medium, respectively. Other risks of bias were considered either low or even trivial.

RQ4) What are the significant early life and environmental factors that are associated with increased risk of IMIDs using data mining techniques?

The factors that were significantly association with a higher risk of IMIDs in Ontario children included “general childhood infection”, “use of antibiotics”, “streptococcus pyogenes”, “respiratory infection”, “gastroenteritis”, “mother's prevalence of any IMID”, and “baby's sex” as risk factors. More specifically, we observed the following effects:

- Use of antibiotics was associated with protectiveness against asthma.
- Having streptococcus pyogenes while the mother had no history of asthma was associated with protectiveness against asthma.
- Born from an immigrant mother with child’s birth weight of less than 3 kg was associated with high risk to develop asthma.
- Mother lived in a rural area during pregnancy* and did not have labour during birth of the child, but the child had respiratory infection during infancy was associated with high risk to develop asthma.

  ** High risk of misclassification bias*
- Having general childhood infection during infancy when the mother had history of asthma was associated with high risk to develop asthma.
- Not having gastroenteritis while born in summertime (July)** was associated with protectiveness against IBD.

  *** Potential risk of confounding*

- Having gastroenteritis while the mother either had a maternal health problem or no labour at birth was associated with high risk to develop IBD.
- Born in an area with less polluted air (PM2.5 level less than 8.9) was associated with protectiveness against IBD, while born in an area with polluted air was associated with high risk to develop IBD.
- Mother with history of SARD was associated with high risk to develop SARDs.
- Use of antibiotics or having streptococcus pyogenes or first respiratory infection at 2 years of again, while the mother had no history of SARD was associated with protectiveness against SARDs.

9.3 Contributions

Briefly, the main contributions of this thesis are:

1. Clarifying the terminology sourced from various multiple fields of science.
2. Designing a new and generic data mining methodology to reliably analyze health administrative data in an unbiased, valid, and sustainable manner.
3. Forming a Technical Preprocessing Guideline with clear sequence and actions, customized for analyzing health administrative data using data mining techniques.
4. Creating a Contextual Preprocessing Guideline to identify and mitigate the risk of applicable sources of bias and confounding in the designed methodology.
5. Building models in three real-world case studies to identify the associations between early life exposures and IMIDs and generate new hypotheses by implementing the newly designed reliable data mining methodology on real health administrative data.

9.4 Limitations

In the process of designing the reliable methodology, we encountered with the following limitations:

- **Narrowing down the type of health data:** Originally, we intended to design a data mining methodology which provides reliable results when analyzing routinely collected health data. It included clinical data, administrative data, and any other type of health data that are gathered routinely. However, we saw that each type of health data has different characteristics that need to be considered during data mining. Therefore, in order to define certain points and directions in the new data mining methodology and to consider the specific sources of bias applicable to each type of health data, we narrowed down the scope of the methodology to only health administrative data.
- **Generalizing the lifelong learning feedback:** As we planned to design a methodology which provides reliable results, we considered a feedback stage which upgraded the models to lifelong learning versions. However, it was clarified that there are many parameters, including the complexity within the data, involved to build usable models when data mining techniques are applied to data. In addition, reliable results do not necessarily mean high-quality models or valid findings. The goal of a reliable methodology is to provide results and evaluations the user can trust. This includes results that the methodology evaluates to be invalid, and the user can trust this evaluation. Therefore, as the resulting models are not necessarily usable enough to benefit from being upgraded to lifelong learning and they might require modifications

in different levels to improve their performance, the feedback stage of the designed methodology was transformed to a more generalized version.

Most other limitations we faced were in the case studies as those were the sections which we dealt with while processing data and implementing methods. These limitations were presented in section 6.6.

9.5 Future Work

The research conducted in this study met all the defined objectives which included the design of a new data mining methodology and implementation of three case studies. However, the work can be continued in various aspects which were out of this study's scope. The potential future works are presented below.

- **Clinical data:** The designed methodology can be investigated to verify which parts, especially in the Contextual Preprocessing Guideline, are applicable to clinical data as well as health administrative data and how it can be modified to be suitable for applying data mining techniques to health clinical data.
- **Unstructured data:** Health administrative data are typically collected in a structured format. However, if the designed methodology is customized for other types of health data, such as clinical data, the required preprocessing steps for unstructured data need to be included. This requires additional investigation on what natural language processing (NLP) techniques may be more suitable for the focused health data.
- **Imputation of high-dimensional data:** Imputing missing values in high-dimensional data is complex, and there are many imputation techniques that can be considered. Further investigations can be done to identify which imputation techniques, including machine learning algorithms (like KNN) or statistical approaches (like multiple

imputation) or even a combination of them, would work efficiently with higher precision on high-dimensional health data.

- **e-Health solution:** If the predictive models built as part of the case studies had higher quality so that the models could be used in production to predict patients, a predictive alerting system on the focused outcomes in a particular health care provider system, such as a clinic or hospital, could have been designed as an electronic health solution resulting from the built models.
- **Lifelong learning on routinely collected data:** The designed methodology is capable of handling routinely collected health administrative data and also upgrading the models to lifelong learning versions. However, it can be expanded to provide more custom directions on how to upgrade the models to lifelong learning to continuously consume routinely collected health administrative data.

REFERENCES

[1] E. I. Benchimol *et al.*, 'The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement', *PLoS Med*, vol. 12, no. 10, p. e1001885, Oct. 2015, doi: 10.1371/journal.pmed.1001885.

[2] R. Lieb, 'Population-Based Study', in *Encyclopedia of Behavioral Medicine*, M. D. Gellman and J. R. Turner, Eds. Springer New York, 2013, pp. 1507–1508. doi: 10.1007/978-1-4419-1005-9_45.

[3] G. S. Cooper, M. L. K. Bynum, and E. C. Somers, 'Recent insights in the epidemiology of autoimmune diseases: Improved prevalence estimates and understanding of clustering of diseases', *J. Autoimmun.*, vol. 33, no. 3–4, pp. 197–207, Nov. 2009, doi: 10.1016/j.jaut.2009.09.008.

[4] S. Agrawal, N. Pearce, and S. Ebrahim, 'Prevalence and risk factors for self-reported asthma in an adult Indian population: a cross-sectional survey', *Int. J. Tuberc. Lung Dis. Off. J. Int. Union Tuberc. Lung Dis.*, vol. 17, no. 2, pp. 275–282, Feb. 2013, doi: 10.5588/ijtld.12.0438.

[5] P. Ranasinghe, R. Jayawardena, and P. Katulanda, 'Diabetes mellitus in South Asia: Scientific evaluation of the research output', *J. Diabetes*, vol. 5, no. 1, pp. 34–42, Mar. 2013, doi: 10.1111/1753-0407.12003.

[6] Y. Ko, R. Butcher, and R. W. Leong, 'Epidemiological studies of migration and environmental risk factors in the inflammatory bowel diseases', *World J. Gastroenterol. WJG*, vol. 20, no. 5, pp. 1238–1247, Feb. 2014, doi: 10.3748/wjg.v20.i5.1238.

[7] E. I. Benchimol *et al.*, 'Asthma, type 1 and type 2 diabetes mellitus, and inflammatory bowel disease amongst South Asian immigrants to Canada and their children: a population-based cohort study', *PloS One*, vol. 10, no. 4, 2015, doi: 10.1371/journal.pone.0123599.

[8] E. I. Benchimol *et al.*, 'Inflammatory Bowel Disease in Immigrants to Canada And Their Children: A Population-Based Cohort Study', *Am. J. Gastroenterol.*, vol. 110, no. 4, pp. 553–563, Apr. 2015, doi: 10.1038/ajg.2015.52.

[9] A. Aujnarain, D. R. Mack, and E. I. Benchimol, 'The Role of the Environment in the Development of Pediatric Inflammatory Bowel Disease', *Curr. Gastroenterol. Rep.*, vol. 15, no. 6, p. 326, Jun. 2013, doi: 10.1007/s11894-013-0326-4.

[10] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Elsevier, 2011.

[11] 'The Technology Review Ten', *MIT Technology Review*, Feb. 2001.

[12] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, 'The KDD Process for Extracting Useful Knowledge from Volumes of Data', *Commun ACM*, vol. 39, no. 11, pp. 27–34, Nov. 1996, doi: 10.1145/240455.240464.

[13] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2006.

[14] J. Natale, 'Leveraging Technology to Revolutionize Canadian Health Care', *Policy: Canadian Politics and Public Policy*, vol. 2, no. 6, pp. 27–30, Dec. 2014.

[15] D. Rudoler, A. Laporte, J. Barnsley, R. H. Glazier, and R. B. Deber, 'Paying for primary care: A cross-sectional analysis of cost and morbidity distributions across primary care payment models in Ontario Canada', *Soc. Sci. Med.*, vol. 124, pp. 18–28, Jan. 2015, doi: 10.1016/j.socscimed.2014.11.001.

[16] E. I. Benchimol *et al.*, ‘Increasing incidence of paediatric inflammatory bowel disease in Ontario, Canada: evidence from health administrative data’, *Gut*, vol. 58, no. 11, pp. 1490–1497, Nov. 2009, doi: 10.1136/gut.2009.188383.
[17] A. R. Hevner, S. T. March, J. Park, and S. Ram, ‘Design Science in Information Systems Research’, *MIS Q.*, vol. 28, no. 1, pp. 75–105, 2004, doi: 10.2307/25148625.
[18] M. H. Tekieh, B. Raahemi, and E. I. Benchimol, ‘Leveraging Applications of Data Mining in Healthcare Using Big Data Analytics’, in *Handbook of Research on Data Science for Effective Healthcare Practice and Administration*, vol. (in-press), IGI Global, 2017.
[19] M. H. Tekieh and B. Raahemi, ‘Importance of Data Mining in Healthcare: A Survey’, in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, New York, NY, USA, 2015, pp. 1057–1062. doi: 10.1145/2808797.2809367.
[20] D. LeSueur, ‘5 Reasons Healthcare Data Is Unique and Difficult to Measure’, *Health Catalyst*, Jun. 12, 2014. https://www.healthcatalyst.com/insights/5-reasons-healthcare-data-is-difficult-to-measure (accessed Mar. 02, 2021).
[21] S. R. Deeny and A. Steventon, ‘Making sense of the shadows: priorities for creating a learning healthcare system based on routinely collected data’, *BMJ Qual. Saf.*, vol. 24, no. 8, pp. 505–515, Aug. 2015, doi: 10.1136/bmjqs-2015-004278.
[22] L. L. Roos, N. C. Nickel, P. S. Romano, and P. Fergusson, ‘Administrative Databases’, in *Wiley StatsRef: Statistics Reference Online*, John Wiley & Sons, Ltd, 2014. Accessed: Mar. 01, 2016. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/9781118445112.stat05227.pub2/abstract
[23] R. C. Dicker, F. Coronado, D. Koo, and R. G. Parrish, *Principles of epidemiology in public health practice; an introduction to applied epidemiology and biostatistics*, Third edition. Centers for Disease Control and Prevention (CDC), 2006.
[24] R. M. Lucas and A. J. McMichael, ‘Association or causation: evaluating links between “environment and disease”.’, *Bull. World Health Organ.*, vol. 83, no. 10, pp. 792–795, Oct. 2005.
[25] K. J. Rothman, S. Greenland, and T. L. Lash, *Modern Epidemiology*. Lippincott Williams & Wilkins, 2008.
[26] M. Porta, *A Dictionary of Epidemiology*, Sixth edition. Oxford University Press, 2014.
[27] R. I. Horwitz and A. R. Feinstein, ‘Exclusion Bias and the False Relationship of Reserpine and Breast Cancer’, *Arch. Intern. Med.*, vol. 145, no. 10, pp. 1873–1875, Oct. 1985, doi: 10.1001/archinte.1985.00360100139023.
[28] B. C. K. Choi and A. W. P. Pak, ‘Bias, Overview’, in *Encyclopedia of Epidemiologic Methods*, Chichester: John Wiley & Sons, 2000, pp. 74–82.
[29] M. J. M. Glesby and D. R. Hoover, ‘Survivor Treatment Selection Bias in Observational Studies: Examples from the AIDS Literature’, *Ann. Intern. Med.*, vol. 124, no. 11, pp. 999–1005, Jun. 1996, doi: 10.7326/0003-4819-124-11-199606010-00008.
[30] J. Neyman, ‘Statistics--Servant of All Sciences’, *Science*, vol. 122, no. 3166, pp. 401–406, 1955.
[31] G. Hill, J. Connelly, R. Hébert, J. Lindsay, and W. Millar, ‘Neyman’s bias re-visited’, *J. Clin. Epidemiol.*, vol. 56, no. 4, pp. 293–296, Apr. 2003, doi: 10.1016/S0895-4356(02)00571-1.

[32] D. L. Sackett, ‘Bias in Analytic Research’, in *The Case-Control Study Consensus and Controversy*, M. A. Ibrahim, Ed. Pergamon, 1979, pp. 51–63. doi: 10.1016/B978-0-08-024907-0.50013-4.
[33] G. S. Fletcher, *Clinical Epidemiology: The Essentials*. Lippincott Williams & Wilkins, 2019.
[34] D. J. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*. MIT Press, 2001.
[35] I. Yoo *et al.*, ‘Data Mining in Healthcare and Biomedicine: A Survey of the Literature’, *J. Med. Syst.*, vol. 36, no. 4, pp. 2431–2448, May 2011, doi: 10.1007/s10916-011-9710-5.
[36] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, ‘From Data Mining to Knowledge Discovery in Databases’, *Commun ACM*, vol. 39, no. 11, pp. 24–26, 1996.
[37] S. Velickov and D. Solomatine, ‘Predictive Data Mining: Practical Examples’, Cottbus, Germany, Mar. 2000.
[38] D. Tomar and S. Agarwal, ‘A survey on Data Mining approaches for Healthcare’, *Int. J. Bio-Sci. Bio-Technol.*, vol. 5, no. 5, pp. 241–266, 2013.
[39] H. Hu, J. Li, A. Plank, H. Wang, and G. Daggard, ‘A Comparative Study of Classification Methods for Microarray Data Analysis’, in *Proceedings of the Fifth Australasian Conference on Data Mining and Analystics*, Darlinghurst, Australia, Australia, 2006, vol. 61, pp. 33–37. Accessed: Apr. 22, 2015. [Online]. Available: http://dl.acm.org/citation.cfm?id=1273808.1273813
[40] H. Cataloluk and M. Kesler, ‘A diagnostic software tool for skin diseases with basic and weighted K-NN’, in *2012 International Symposium on Innovations in Intelligent Systems and Applications (INISTA)*, Jul. 2012, pp. 1–4. doi: 10.1109/INISTA.2012.6246999.
[41] R. Potter, ‘Comparison of classification algorithms applied to breast cancer diagnosis and prognosis’, Leipzig, Germany, Jul. 2007, pp. 40–49.
[42] G. A. Beller, ‘The rising cost of health care in the United States: Is it making the United States globally noncompetitive?’, *J. Nucl. Cardiol.*, vol. 15, no. 4, pp. 481–482, Jul. 2008, doi: 10.1016/j.nuclcard.2008.06.001.
[43] D. Bertsimas *et al.*, ‘Algorithmic Prediction of Health-Care Costs’, *Oper. Res.*, vol. 56, no. 6, pp. 1382–1392, Dec. 2008, doi: 10.1287/opre.1080.0619.
[44] M. H. Tekieh, B. Raahemi, and S. A. Izad Shenas, ‘Analysing healthcare coverage with data mining techniques’, *Int. J. Soc. Syst. Sci.*, vol. 7, no. 3, pp. 198–221, 2015.
[45] D. J. Hand, ‘Data Mining: Statistics and More?’, *Am. Stat.*, vol. 52, no. 2, pp. 112–118, May 1998, doi: 10.1080/00031305.1998.10480549.
[46] D. J. Hand, ‘Statistics and Data Mining: Intersecting Disciplines’, *SIGKDD Explor Newsl*, vol. 1, no. 1, pp. 16–19, Jun. 1999, doi: 10.1145/846170.846171.
[47] ‘Healthways Heads Off Increased Costs with SAS’. SAS, 2009. [Online]. Available: http://www.sas.com/success/pdf/healthways.pdf
[48] Y. Zhang, S. Fong, S. Fiaidhi, and S. Mohammed, ‘Real-time clinical decision support system with data stream mining’, *J. Biomed. Biotechnol.*, vol. 2012, p. 8, 2012.
[49] D. Delen, G. Walker, and A. Kadam, ‘Predicting breast cancer survivability: a comparison of three data mining methods’, *Artif. Intell. Med.*, vol. 34, no. 2, pp. 113–127, Jun. 2005, doi: 10.1016/j.artmed.2004.07.002.
[50] S. Shah, A. Kusiak, and B. Dixon, ‘Data mining in predicting survival of kidney dialysis patients’, in *Proceedings of Photonics West—Bios 2003*, Belingham, 2003, vol. 4949, pp. 73–79. doi: 10.1117/12.476387.

[51] T. Haferlach *et al.*, 'Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: Report from the international microarray innovations in leukemia study group', *J. Clin. Oncol.*, vol. 28, no. 15, pp. 2529–2537, 2010.
[52] R. Salazar *et al.*, 'Gene expression signature to improve prognosis prediction of stage ii and iii colorectal cancer', *J. Clin. Oncol.*, vol. 29, no. 1, pp. 17–24, 2011.
[53] T. R. Golub *et al.*, 'Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring', *Science*, vol. 286, no. 5439, pp. 531–537, Oct. 1999, doi: 10.1126/science.286.5439.531.
[54] 'First Things First—Highmark makes healthcare-fraud prevention top priority with SAS'. SAS, 2006. [Online]. Available: http://www.sas.com/success/pdf/highmarkfraud.pdf
[55] M.-H. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki, and D. K. Grunwell, 'Health big data analytics: current perspectives, challenges and potential solutions', *Int. J. Big Data Intell.*, vol. 1, no. 1/2, pp. 114–126, 2014.
[56] R. D. Canlas Jr, 'Data mining in healthcare: Current Applications and Issues', Carnegie Mellon University, Australia, 2009.
[57] F. Hosseinkhah, H. Ashktorab, R. Veen, and M. M. Owrang O., 'Challenges in Data Mining on Medical Databases', *IGI Glob.*, pp. 502–511, 2009.
[58] S. Hoffman and A. Podgurski, 'Big Bad Data: Law, Public Health, and Biomedical Databases', *J. Law. Med. Ethics*, vol. 41, pp. 56–60, Mar. 2013, doi: 10.1111/jlme.12040.
[59] C. Violán *et al.*, 'Comparison of the information provided by electronic health records data and a population health survey to estimate prevalence of selected health conditions and multimorbidity', *BMC Public Health*, vol. 13, no. 1, p. 251, Mar. 2013, doi: 10.1186/1471-2458-13-251.
[60] K. El Emam, *Guide to the De-Identification of Personal Health Information*. CRC Press, 2013.
[61] 'The Four V's of Big Data', *IBM Big Data & Analytics Hub*, 2013. http://www.ibmbigdatahub.com/infographic/four-vs-big-data (accessed May 31, 2016).
[62] J. M. Hardin and D. C. Chhieng, 'Data Mining and Clinical Decision Support Systems', in *Clinical Decision Support Systems*, E. S. B. E. FHIMSS FACMI, Ed. Springer New York, 2007, pp. 44–63. doi: 10.1007/978-0-387-38319-4_3.
[63] S. Bandyopadhyay *et al.*, 'Data mining for censored time-to-event data: a Bayesian network model for predicting cardiovascular risk from electronic health record data', *Data Min. Knowl. Discov.*, vol. 29, no. 4, pp. 1033–1069, Jul. 2015, doi: 10.1007/s10618-014-0386-6.
[64] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, 'Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients', *Health Aff. (Millwood)*, vol. 33, no. 7, pp. 1123–1131, Jul. 2014, doi: 10.1377/hlthaff.2014.0041.
[65] J. F. Easton, C. R. Stephens, and M. Angelova, 'Risk factors and prediction of very short term versus short/intermediate term post-stroke mortality: A data mining approach', *Comput. Biol. Med.*, vol. 54, pp. 199–210, Nov. 2014, doi: 10.1016/j.compbiomed.2014.09.003.
[66] D. Raju, X. Su, P. A. Patrician, L. A. Loan, and M. S. McCarthy, 'Exploring factors associated with pressure ulcers: A data mining approach', *Int. J. Nurs. Stud.*, vol. 52, no. 1, pp. 102–111, Jan. 2015, doi: 10.1016/j.ijnurstu.2014.08.002.

[67] M. Burroni *et al.*, 'Melanoma Computer-Aided Diagnosis: Reliability and Feasibility Study', *Am. Assoc. Cancer Res.*, vol. 10, no. 6, pp. 1881–1886, Mar. 2004, doi: 10.1158/1078-0432.CCR-03-0039.
[68] H.-J. Chen, R. Chen, M. Yang, G.-J. Teng, and E. H. Herskovits, 'Identification of Minimal Hepatic Encephalopathy in Patients with Cirrhosis Based on White Matter Imaging and Bayesian Data Mining', *Am. J. Neuroradiol.*, vol. 36, no. 3, pp. 481–487, Mar. 2015, doi: 10.3174/ajnr.A4146.
[69] S. Joo, Y. S. Yang, W. K. Moon, and H. C. Kim, 'Computer-aided diagnosis of solid breast nodules: use of an artificial neural network based on multiple sonographic features', *IEEE Trans. Med. Imaging*, vol. 23, no. 10, pp. 1292–1300, Oct. 2004, doi: 10.1109/TMI.2004.834617.
[70] B. B. Zellner, S. D. Rand, R. Prost, H. Krouwer, and V. K. Chetty, 'A cost-minimizing diagnostic methodology for discrimination between neoplastic and non-neoplastic brain lesions: utilizing a genetic algorithm scientific reports', *Acad. Radiol.*, vol. 11, no. 2, pp. 169–177, Feb. 2004, doi: 10.1016/S1076-6332(03)00654-8.
[71] S. Ram, W. Zhang, M. Williams, and Y. Pengetnze, 'Predicting Asthma-Related Emergency Department Visits Using Big Data', *IEEE J. Biomed. Health Inform.*, vol. 19, no. 4, pp. 1216–1223, Jul. 2015, doi: 10.1109/JBHI.2015.2404829.
[72] G. S. Campos, A. C. Bandeira, and S. I. Sardi, 'Zika Virus Outbreak, Bahia, Brazil', *Emerg. Infect. Dis.*, vol. 21, no. 10, pp. 1885–1886, Oct. 2015, doi: 10.3201/eid2110.150847.
[73] WHO, 'Chronic respiratory diseases: Asthma', *World Health Organization*, 2017. http://www.who.int/respiratory/asthma/en/ (accessed Jan. 22, 2017).
[74] N. Bock, M. Scholz, G. Piller, and K. Böhm, 'Capture and Analysis of Sensor Data for Asthma Patients (Research-in-Progress)', presented at the Twenty-Fourth European Conference on Information Systems (ECIS), Istanbul,Turkey, 2016.
[75] C. M. Perou *et al.*, 'Molecular portraits of human breast tumours', *Nature*, vol. 406, no. 6797, pp. 747–752, Aug. 2000, doi: 10.1038/35021093.
[76] R. J. Robison, 'How big is the human genome? In megabytes, not base pairs.', *Medium*, Jan. 06, 2014. https://medium.com/precision-medicine/how-big-is-the-human-genome-e90caa3409b0 (accessed Sep. 13, 2016).
[77] J. M. Geraci *et al.*, 'Mortality after non-cardiac surgery: Prediction from administrative versus clinical data', *J. Investig. Med.*, vol. 48, no. 2, Mar. 2000.
[78] S. Hajian, F. Bonchi, and C. Castillo, 'Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining', in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2016, pp. 2125–2126. doi: 10.1145/2939672.2945386.
[79] T. Razzaghi, O. Roderick, I. Safro, and N. Marko, 'Fast imbalanced classification of healthcare data with missing values', in *2015 18th International Conference on Information Fusion (Fusion)*, Jul. 2015, pp. 774–781.
[80] T. Razzaghi, O. Roderick, I. Safro, and N. Marko, 'Multilevel Weighted Support Vector Machine for Classification on Healthcare Data with Missing Values', *PLoS ONE*, vol. 11, no. 5, May 2016, Accessed: Feb. 28, 2017. [Online]. Available: http://go.galegroup.com/ps/i.do?p=AONE&sw=w&issn=19326203&v=2.1&it=r&id=GALE%7CA453360149&sid=googleScholar&linkaccess=abs
[81] J. J. Tepas, 'Data mining: childhood injury control and beyond', *J. Trauma*, vol. 67, no. 2 Suppl, pp. S108-110, Aug. 2009, doi: 10.1097/TA.0b013e3181af0ad7.

[82] W. Zhou and H. Xiong, ‘Efficient Discovery of Confounders in Large Data Sets’, in *2009 Ninth IEEE International Conference on Data Mining*, Dec. 2009, pp. 647–656. doi: 10.1109/ICDM.2009.77.
[83] J. C. Ferrão, M. D. Oliveira, F. Janela, and H. M. G. Martins, ‘Preprocessing structured clinical data for predictive modeling and decision support’, *Appl. Clin. Inform.*, vol. 7, no. 4, pp. 1135–1153, 2016, doi: 10.4338/ACI-2016-03-SOA-0035.
[84] C.-H. Lee, J. C.-Y. Chen, and V. S. Tseng, ‘A novel data mining mechanism considering bio-signal and environmental data with applications on asthma monitoring’, *Comput. Methods Programs Biomed.*, vol. 101, no. 1, pp. 44–61, Jan. 2011, doi: 10.1016/j.cmpb.2010.04.016.
[85] V. S. Tseng, C.-H. Lee, and J. Chia-Yu Chen, ‘An Integrated Data Mining System for Patient Monitoring with Applications on Asthma Care’, in *21st IEEE International Symposium on Computer-Based Medical Systems, 2008. CBMS ’08*, Jun. 2008, pp. 290–292. doi: 10.1109/CBMS.2008.111.
[86] B. J. Bereznicki, G. M. Peterson, S. L. Jackson, E. H. Walters, Fitzmaurice, and P. R. Gee, ‘Data-mining of medication records to improve asthma management’, *Med. J. Aust.*, vol. 189, no. 1, pp. 21–25, Jul. 2008.
[87] M. Gironi *et al.*, ‘A novel data mining system points out hidden relationships between immunological markers in multiple sclerosis’, *Immun. Ageing*, vol. 10, no. 1, pp. 1–7, Jan. 2013, doi: 10.1186/1742-4933-10-1.
[88] P. Guo, Q. Zhang, Z. Zhu, Z. Huang, and K. Li, ‘Mining Gene Expression Data of Multiple Sclerosis’, *PLoS ONE*, vol. 9, no. 6, p. e100052, Jun. 2014, doi: 10.1371/journal.pone.0100052.
[89] M. Marinov, A. S. M. Mosa, I. Yoo, and S. A. Boren, ‘Data-Mining Technologies for Diabetes: A Systematic Review’, *J. Diabetes Sci. Technol.*, vol. 5, no. 6, pp. 1549–1556, Nov. 2011, doi: 10.1177/193229681100500631.
[90] I. C. Gerling, S. Singh, N. I. Lenchik, D. R. Marshall, and J. Wu, ‘New Data Analysis and Mining Approaches Identify Unique Proteome and Transcriptome Markers of Susceptibility to Autoimmune Diabetes’, *Mol. Cell. Proteomics*, vol. 5, no. 2, pp. 293–305, Feb. 2006, doi: 10.1074/mcp.M500197-MCP200.
[91] R. Bhramaramba, A. R. Allam, V. V. Kumar, and G. R. Sridhar, ‘Application of data mining techniques on diabetes related proteins’, *Int. J. Diabetes Dev. Ctries.*, vol. 31, no. 1, pp. 22–25, Jan. 2011, doi: 10.1007/s13410-010-0001-3.
[92] A. A. Aljumah, M. K. Siddiqui, and M. G. Ahamad, ‘Application of Classification based Data Mining Technique in Diabetes Care’, *J. Appl. Sci.*, vol. 13, pp. 416–422, Mar. 2013, doi: 10.3923/jas.2013.416.422.
[93] B. R. Shah and L. L. Lipscombe, ‘Clinical Diabetes Research Using Data Mining: A Canadian Perspective’, *Can. J. Diabetes*, vol. 39, no. 3, pp. 235–238, Jun. 2015, doi: 10.1016/j.jcjd.2015.02.005.
[94] X.-H. Meng, Y.-X. Huang, D.-P. Rao, Q. Zhang, and Q. Liu, ‘Comparison of three data mining models for predicting diabetes or prediabetes by risk factors’, *Kaohsiung J. Med. Sci.*, vol. 29, no. 2, pp. 93–99, Feb. 2013, doi: 10.1016/j.kjms.2012.08.016.
[95] L. Tapak, H. Mahjub, O. Hamidi, and J. Poorolajal, ‘Real-Data Comparison of Data Mining Methods in Prediction of Diabetes in Iran’, *Healthc. Inform. Res.*, vol. 19, no. 3, pp. 177–185, Sep. 2013, doi: 10.4258/hir.2013.19.3.177.

[96] D. Gregori *et al.*, 'Using Data Mining Techniques in Monitoring Diabetes Care. The Simpler the Better?', *J. Med. Syst.*, vol. 35, no. 2, pp. 277–281, Sep. 2009, doi: 10.1007/s10916-009-9363-9.
[97] V. Vaishnavi, B. Kuechler, and S. Petter, 'Design Science Research in Information Systems'. Association for Information Systems, 2004. [Online]. Available: http://desrist.org/design-research-in-information-systems/
[98] J. E. V. Aken, 'Management Research as a Design Science: Articulating the Research Products of Mode 2 Knowledge Production in Management', *Br. J. Manag.*, vol. 16, no. 1, pp. 19–36, 2005, doi: https://doi.org/10.1111/j.1467-8551.2005.00437.x.
[99] P. Chapman *et al.*, 'CRISP-DM 1.0: Step-by-Step Data Mining Guide', SPSS inc, Aug. 2000.
[100] 'Poll: What main methodology are you using for your analytics, data mining, or data science projects?', *KDNuggets*, Oct. 2014. https://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html (accessed Mar. 11, 2021).
[101] G. Mariscal, Ó. Marbán, and C. Fernández, 'A survey of data mining and knowledge discovery process models and methodologies', *Knowl. Eng. Rev.*, vol. 25, no. 2, pp. 137–166, Jun. 2010, doi: http://dx.doi.org.proxy.bib.uottawa.ca/10.1017/S0269888910000032.
[102] A. Bronshtein, 'Train/Test Split and Cross Validation in Python', *Medium*, Mar. 24, 2020. https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6 (accessed Sep. 26, 2021).
[103] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, 1st edition. Boston: Pearson, 2005.
[104] N. A. Molodecky, R. Panaccione, S. Ghosh, H. W. Barkema, G. G. Kaplan, and Alberta Inflammatory Bowel Disease Consortium, 'Challenges associated with identifying the environmental determinants of the inflammatory bowel diseases', *Inflamm. Bowel Dis.*, vol. 17, no. 8, pp. 1792–1799, Aug. 2011, doi: 10.1002/ibd.21511.
[105] S. Thrun, 'Lifelong Learning Algorithms', in *Learning to Learn*, Springer, Boston, MA, 1998, pp. 181–209. doi: 10.1007/978-1-4615-5529-2_8.
[106] C. van Walraven and P. Austin, 'Administrative database research has unique characteristics that can risk biased results', *J. Clin. Epidemiol.*, vol. 65, no. 2, pp. 126–131, Feb. 2012, doi: 10.1016/j.jclinepi.2011.08.002.
[107] C. van Walraven, C. Bennett, and A. J. Forster, 'Administrative database research infrequently used validated diagnostic or procedural codes', *J. Clin. Epidemiol.*, vol. 64, no. 10, pp. 1054–1059, Oct. 2011, doi: 10.1016/j.jclinepi.2011.01.001.
[108] T. Burgmann *et al.*, 'The Manitoba Inflammatory Bowel Disease Cohort Study: prolonged symptoms before diagnosis--how much is irritable bowel syndrome?', *Clin. Gastroenterol. Hepatol. Off. Clin. Pract. J. Am. Gastroenterol. Assoc.*, vol. 4, no. 5, pp. 614–620, May 2006, doi: 10.1016/j.cgh.2006.03.003.
[109] M. E. Kuenzig *et al.*, 'Asthma is not associated with the need for surgery in Crohn's disease when controlling for smoking status: a population-based cohort study', *Clin. Epidemiol.*, vol. 10, pp. 831–840, 2018, doi: 10.2147/CLEP.S156772.
[110] T. A. Stukel, E. S. Fisher, D. E. Wennberg, D. A. Alter, D. J. Gottlieb, and M. J. Vermeulen, 'Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and

instrumental variable methods', *JAMA*, vol. 297, no. 3, pp. 278–285, Jan. 2007, doi: 10.1001/jama.297.3.278.
[111] A. Guttmann *et al.*, 'Validation of a health administrative data algorithm for assessing the epidemiology of diabetes in Canadian children', *Pediatr. Diabetes*, vol. 11, no. 2, pp. 122–128, Mar. 2010, doi: 10.1111/j.1399-5448.2009.00539.x.
[112] T. To, A. Gershon, C. Wang, S. Dell, and L. Cicutto, 'Persistence and remission in childhood asthma: a population-based asthma birth cohort study', *Arch. Pediatr. Adolesc. Med.*, vol. 161, no. 12, pp. 1197–1204, Dec. 2007, doi: 10.1001/archpedi.161.12.1197.
[113] E. I. Benchimol *et al.*, 'Incidence, Outcomes, and Health Services Burden of Very Early Onset Inflammatory Bowel Disease', *Gastroenterology*, vol. 147, no. 4, pp. 803-813.e7, Oct. 2014, doi: 10.1053/j.gastro.2014.06.023.
[114] D. Rigante, A. Bosco, and S. Esposito, 'The Etiology of Juvenile Idiopathic Arthritis', *Clin. Rev. Allergy Immunol.*, vol. 49, no. 2, pp. 253–261, Oct. 2015, doi: 10.1007/s12016-014-8460-9.
[115] H. Peng and W. Hagopian, 'Environmental factors in the development of Type 1 diabetes', *Rev. Endocr. Metab. Disord.*, vol. 7, no. 3, pp. 149–162, Sep. 2006, doi: 10.1007/s11154-006-9024-y.
[116] S. Dick *et al.*, 'A systematic review of associations between environmental exposures and development of asthma in children aged up to 9 years', *BMJ Open*, vol. 4, no. 11, p. e006554, Nov. 2014, doi: 10.1136/bmjopen-2014-006554.
[117] R. Beasley, A. Semprini, and E. A. Mitchell, 'Risk factors for asthma: is prevention possible?', *The Lancet*, vol. 386, no. 9998, pp. 1075–1085, Sep. 2015, doi: 10.1016/S0140-6736(15)00156-7.
[118] D. N. Krementsov and C. Teuscher, 'Environmental factors acting during development to influence MS risk: insights from animal studies', *Mult. Scler. J.*, vol. 19, no. 13, pp. 1684–1689, Nov. 2013, doi: 10.1177/1352458513506954.
[119] 'Better Outcomes Registry and Network (BORN) - Data Library Details', *ICES Data Dictionary*. (secure access) (accessed Nov. 15, 2021).
[120] 'Ontario Crohn's and Colitis Cohort dataset (OCCC) - Data Library Details', *ICES Data Dictionary*. (secure access) (accessed Nov. 15, 2021).
[121] 'Ontario Asthma (ASTHMA) - Data Library Details', *ICES Data Dictionary*. (secure access) (accessed Nov. 15, 2021).
[122] 'Ontario Diabetes Dataset (ODD) - Data Library Details', *ICES Data Dictionary*. (secure access) (accessed Nov. 15, 2021).
[123] N. J. Shiff *et al.*, 'The prevalence of systemic autoimmune rheumatic diseases in Canadian pediatric populations: administrative database estimates', *Rheumatol. Int.*, vol. 35, no. 3, pp. 569–573, Mar. 2015, doi: 10.1007/s00296-014-3136-6.
[124] J. Widdifield *et al.*, 'Development and validation of an administrative data algorithm to estimate the disease burden and epidemiology of multiple sclerosis in Ontario, Canada', *Mult. Scler. Houndmills Basingstoke Engl.*, vol. 21, no. 8, pp. 1045–1054, Jul. 2015, doi: 10.1177/1352458514556303.
[125] T. To *et al.*, 'Case verification of children with asthma in Ontario', *Pediatr. Allergy Immunol.*, vol. 17, no. 1, pp. 69–76, 2006, doi: 10.1111/j.1399-3038.2005.00346.x.
[126] A. Pisesky *et al.*, 'Incidence of Hospitalization for Respiratory Syncytial Virus Infection amongst Children in Ontario, Canada: A Population-Based Study Using

Validated Health Administrative Data', *PLOS ONE*, vol. 11, no. 3, p. e0150416, Mar. 2016, doi: 10.1371/journal.pone.0150416.

[127] 'Number of births in Ontario, Canada 2021', *Statista*. https://www.statista.com/statistics/578584/number-of-births-in-ontario-canada/ (accessed Nov. 29, 2021).

APPENDICES

Appendix I. IMIDs Early Life and Environmental Factors

Potential early life and environmental factors for immune-mediated inflammatory diseases retrieved from literature are as following:

Category	Factor	Observed in literature	Effect	Exists in ICES datasets
Perinatal	Maternal smoking	IBD, Asthma	Risk/Protective	✓
	Paternal smoking	Asthma	Risk/Protective	
	Child smoking	Asthma	Risk/Protective	
	Maternal age	IBD, T1DM	Risk/Protective	✓
	Maternal asthma	Asthma	Risk/Protective	✓
	Maternal weight gain	T1DM	Risk/Protective	✓
	Amniocentesis	T1DM	Risk/Protective	✓
	Pre-eclampsia	T1DM	Risk/Protective	✓
	Jaundice	T1DM	Risk/Protective	✓
	Complicated delivery	T1DM	Risk/Protective	✓
	C-section	IBD, Asthma, T1DM	Risk/Protective	✓
	Late premature delivery	Asthma	Risk/Protective	✓
	Firstborn child	T1DM	Risk/Protective	✓
	Low birth weight	Asthma, T1DM	Risk/Protective	✓
	Fast infant weight gain	Asthma, T1DM	Risk/Protective	
	Short birth length	T1DM	Risk/Protective	
Home	Bedroom sharing	IBD	Risk/Protective	
	Bedroom heating	Asthma	Risk/Protective	
	Bedroom painting/redecoration	Asthma	Risk/Protective	
	Family size	IBD	Risk/Protective	✓
	Household crowding	IBD	Risk/Protective	✓
	Cat exposure/pet	IBD, Asthma	Risk/Protective	
	Dog exposure/pet	Asthma	Risk/Protective	
	Personal towel	IBD	Risk/Protective	✓
	Damp housing	Asthma	Risk/Protective	
	Mouse allergen	Asthma	Risk/Protective	
	Feather quilt	Asthma	Risk/Protective	

	Synthetic bedding items	Asthma	Risk/Protective	
	House dust mite	Asthma	Risk/Protective	
	Microbial products	MS	Risk/Protective	
	Cockroach	Asthma	No	
Urban/Rural	Air pollution	IBD, Asthma, T1DM	Risk/Protective	✓
	Traffic-related particles	Asthma	Risk/Protective	✓
	Living in farm	IBD	Risk/Protective	✓
	Urbanisation	Asthma	Risk/Protective	✓
	Birthday during fungal spore season	Asthma	Risk/Protective	✓
	Grass pollen exposure	Asthma	Risk/Protective	✓
	Sunlight exposure/latitude	MS	Risk/Protective	✓
Diet (Mother/ Child)	Animal meats	IBD	Risk/Protective	
	Breastfeeding	IBD, Asthma, T1DM	Risk/Protective	✓
	Cow's milk formula	Asthma	Risk/Protective	✓
	Fats	IBD	Risk/Protective	
	Sweets	IBD	Risk/Protective	
	Fruits/vegetables	IBD, Asthma	Risk/Protective	
	Mediterranean diet	Asthma	Risk/Protective	
	Western diet	Asthma	Risk/Protective	
	Fast food	Asthma	Risk/Protective	
	Fish consumption	Asthma	Risk/Protective	
	Trans- fatty acids	Asthma	Risk/Protective	
	Salt	Asthma	Risk/Protective	
	Peanuts and tree nuts	Asthma	Risk/Protective	
	Chocolate	Asthma	Risk/Protective	
	Vitamins A/D/E	Asthma	Risk/Protective	
	Vitamin D3	MS	Risk/Protective	
	Wheat	Asthma	Risk/Protective	
	Nicotinamide	T1DM	Risk/Protective	
	Zinc	T1DM	Risk/Protective	
	Vitamins C/D/E	T1DM	Risk/Protective	
	Early introduction of cereals, potatoes/carrots, fruit/berries, cow's milk, N-nitroso compounds, and increased calories	T1DM	Risk/Protective	
	Fish oil	Asthma	No	
	Butter/margarine	Asthma	No	
	Organic food	Asthma	No	
	Dietary anti oxidant	Asthma	No	
	General childhood infection	IBD, Asthma	Risk/Protective	✓

Infection/ Illness/ Pathogens	Gastroenteritis	IBD	Risk/Protective	✓
	Respiratory infection	Asthma	Risk/Protective	✓
	Rhinitis	Asthma	Risk/Protective	✓
	Use of antibiotics	IBD, Asthma	Risk/Protective	✓
	Salmonella	IBD	Risk/Protective	✓
	Pertussis	Asthma	Risk/Protective	✓
	Atopy	Asthma	Risk/Protective	✓
	Virus positive wheezing illness	Asthma	Risk/Protective	✓
	Epstein-barr virus infection	MS, JIA, T1DM	Risk/Protective	
	Gamma-herper virus infection	MS	Risk/Protective	✓
	Parvovirus B19	JIA, T1DM	Risk/Protective	✓
	Enteric bacteria	JIA	Risk/Protective	
	Chlamydia trachomatis bacteria	JIA	Risk/Protective	✓
	Chlamydophila pneumoniae bacteria	JIA	Risk/Protective	✓
	Bartonella henselae bacteria	JIA	Risk/Protective	✓
	Mycoplasma pneumoniae bacteria	JIA, T1DM	Risk/Protective	✓
	Streptococcus pyogenes bacteria	JIA, T1DM	Risk/Protective	✓
	Congenital rubella infection	T1DM	Risk/Protective	✓
	Coxsackie B virus	T1DM	Risk/Protective	✓
	Mumps	T1DM	Risk/Protective	✓
	Echovirus	T1DM	Risk/Protective	✓
	Cytomegalovirus	T1DM	Risk/Protective	✓
	Rhinovirus	Asthma	Risk/Protective	✓
	Retrovirus	T1DM	Risk/Protective	✓
	Rotavirus	T1DM	Risk/Protective	✓
	Mycobacterium tuberculosis bacteria	T1DM	Risk/Protective	✓
	Pseudomonas aeruginosa bacteria	T1DM	Risk/Protective	✓
	Yersinia enterocolitica bacteria	T1DM	Risk/Protective	✓
	Mycotoxin bafilomycin A1	T1DM	Risk/Protective	
Operation/ Medication	Appendectomy	IBD	Risk/Protective	✓
	Paracetamol	Asthma	Risk/Protective	
	Gastro-oesophageal reflux treatment	Asthma	Risk/Protective	✓
	Opiates	Asthma	Risk/Protective	✓
	Thyroid supplements	Asthma	Risk/Protective	✓
	Beta agonists	Asthma	Risk/Protective	
Second-Hand Smoke	Passive smoking exposure	IBD, T1DM	Risk/Protective	✓
	Antenatal exposure	Asthma	Risk/Protective	✓
	Postnatal exposure	Asthma	Risk/Protective	✓

Domestic Combustion	Fine particulates	Asthma	Risk/Protective	✓
	Detectable Sulfur Dioxide	Asthma	Risk/Protective	
	Biomass	Asthma	Risk/Protective	
	Gas cooking	Asthma	No	
	Incense	Asthma	No	
Inhaled Chemicals	Volatile organic compound	Asthma	Risk/Protective	
	Chlorinated swimming pools	Asthma	Risk/Protective	
	Other chemicals	Asthma	Risk/Protective	
Other Exposures	Dietary dioxins and polychlorinated biphenyl	Asthma	Risk/Protective	
	Bisphenol A exposure	Asthma	Risk/Protective	
	Bisphenol A exposure	MS	No	
	Increase in dichlorodiphenyltrichloroethane metabolite	Asthma	Risk/Protective	
	Increasing electromagnetic exposure	Asthma	Risk/Protective	
	Occupational exposure	Asthma	Risk/Protective	
	Moulds	Asthma	Risk/Protective	
Demographics/ Lifestyle	Sex	Asthma, MS	Risk/Protective	✓
	Month of birth - due to sunlight exposure	MS	Risk/Protective	✓
	Family history	Asthma	Risk/Protective	✓
	High-income lifestyle	Asthma, IBD	Risk/Protective	✓
	High BMI	Asthma	Risk/Protective	✓
	Sedentary behaviour	Asthma	Risk/Protective	
Psychological	Stress	Asthma, MS, T1DM	Risk/Protective	
	Stress	IBD	No	
	Stressful life events - divorce, disaster, etc.	T1DM	Risk/Protective	
	Anxiety	IBD	No	✓
	Depression	IBD	No	✓
Immigration	Second generation immigrant	IBD	Risk/Protective	✓

Appendix II. University of Ottawa's REB Approval

File Number: 12-16-11 **Date (mm/dd/yyyy):** 12/19/2016

Université d'Ottawa University of Ottawa
Bureau d'éthique et d'intégrité de la recherche Office of Research Ethics and Integrity

Ethics Approval Notice

Social Science and Humanities REB

Principal Investigator / Supervisor / Co-investigator(s) / Student(s)

First Name	Last Name	Affiliation	Role
Bijan	Raahemi	Telfer School of Management	Supervisor
Mohammad Hossein	Tekieh	Telfer School of Management	Student Researcher

File Number: 12-16-11

Type of Project: Doctoral thesis – Secondary use of data

Title: Mining of population-based routinely collected health data to determine risk factors associated with pediatric morbidity in Ontario, Canada

Approval Date (mm/dd/yyyy)	Expiry Date (mm/dd/yyyy)	Approval Type
12/19/2016	12/18/2017	Approval

Special Conditions / Comments:

550, rue Cumberland, pièce 154 550 Cumberland Street, room 154
Ottawa (Ontario) K1N 6N5 Canada Ottawa, Ontario K1N 6N5 Canada
(613) 562-5387 • Téléc./Fax (613) 562-5338
www.recherche.uottawa.ca/deontologie/ www.research.uottawa.ca/ethics/

Université d'Ottawa University of Ottawa
Bureau d'éthique et d'intégrité de la recherche Office of Research Ethics and Integrity

This is to confirm that the University of Ottawa Research Ethics Board identified above, which operates in accordance with the Tri-Council Policy Statement (2010) and other applicable laws and regulations in Ontario, has examined and approved the ethics application for the above named research project. Ethics approval is valid for the period indicated above and subject to the conditions listed in the section entitled "Special Conditions / Comments".

During the course of the project, the protocol may not be modified without prior written approval from the REB except when necessary to remove participants from immediate endangerment or when the modification(s) pertain to only administrative or logistical components of the project (e.g., change of telephone number). Investigators must also promptly alert the REB of any changes which increase the risk to participant(s), any changes which considerably affect the conduct of the project, all unanticipated and harmful events that occur, and new information that may negatively affect the conduct of the project and safety of the participant(s). Modifications to the project, including consent and recruitment documentation, should be submitted to the Ethics Office for approval using the "Modification to research project" form available at: http://research.uottawa.ca/ethics/submissions-and-reviews.

Please submit an annual report to the Ethics Office four weeks before the above-referenced expiry date to request a renewal of this ethics approval. To close the file, a final report must be submitted. These documents can be found at: http://research.uottawa.ca/ethics/submissions-and-reviews.

If you have any questions, please do not hesitate to contact the Ethics Office at extension 5387 or by e-mail at: ethics@uOttawa.ca.

Riana Marcotte
Protocol Officer for Ethics in Research
For Barbara Graves, Chair of the Social Sciences and Humanities REB

550, rue Cumberland, pièce 154 550 Cumberland Street, room 154
Ottawa (Ontario) K1N 6N5 Canada Ottawa, Ontario K1N 6N5 Canada
(613) 562-5387 • Téléc./Fax (613) 562-5338
www.recherche.uottawa.ca/deontologie/ www.research.uottawa.ca/ethics/

Appendix III. Linked Data Libraries for Case Studies

The profile of newborns and their mothers were linked from BORN-Niday dataset to the datasets of data libraries listed below located at ICES to conduct the required investigations in the defined case studies. We have provided preliminary information on their collection and validation procedures too.

- **ASTHMA** *(fiscal 1993/94 to 2016/17)*: The Ontario Asthma Database contains all Ontario asthma patients identified since 1991. The datasets in ASTHMA are created based on OHIP, DAD, and SDS databases, and only include individuals eligible to health coverage in Ontario. Also, a person has to meet the case definition criteria in order to be included in this database; therefore, they could have been diagnosed many years prior to the data availability. The data have been validated in past studies [125], and the sensitivity and specificity for individuals under 18 years of age were 89% and 72%, respectively.
- **CIC** *(fiscal 1985/86 to 2016/17)*: The Ontario portion of Immigration, Refugees and Citizenship Canada (IRCC)'s Permanent Resident Database is available for use at ICES. This includes immigration application records for people who initially applied to land in Ontario. Records date from 1985. The data contain permanent residents' demographic information such as country of citizenship, level of education, mother tongue, and landing date. Note that there are a small number of duplicate IKNs in this data; some of these may be real, in the sense that the person landed in Ontario more than once, while some may be data errors. Some main gaps when this data are linked with other ICES data are: 1) some new immigrants who are currently residing in Ontario but originally landed in another province will not be captured, 2) new

immigrants who landed in Ontario and immediately moved to another province will not have an Ontario health card.

- **DAD** *(fiscal 1988/89 to 2016/17)*: The DAD library consists of two types of data: Discharge Abstract Database (DAD), and Hospital Medical Records Institute (HMRI). These datasets contain patient-level data for acute, rehab, chronic, and day surgery institutions in Ontario. DAD is maintained by the Canadian Institute for Health Information (CIHI). After each patient is discharged from hospital, a medical records coder at the hospital draws up an abstract from the chart, compiling administrative and clinical data on that particular stay. One abstract is completed for each separation (discharge, transfer, death, or stillbirth) from the hospital. The abstract is filled out according to a standard set of instructions. Hospitals forward records to CIHI in one-month batches after (almost) all discharges for a given month are processed. There is usually a 3- or 4-months lag time between the end of the month and the time that CIHI receives the data. CIHI receives data directly from participating hospitals (about 85% of all hospital inpatient discharges in Canada). This is a Canadian total of 4.3 million records annually which typically contains around 1.4 million records in Ontario.
- **MOMBABY** *(fiscal 1988/89 to 2016/17)*: This MOMBABY dataset links the DAD inpatient admission records of delivering mothers and their newborns. The sensitivity and specificity of the algorithm linking mother and newborn records by matching the institutions they were admitted to, their postal codes, and their admission/discharge dates are 0.9613 and 0.9924 for years from 2002/03 to 2009/10 respectively.
- **NACRS** *(fiscal 2000/01 to 2016/17)*: The National Ambulatory Care Reporting System (NACRS) is a data collection tool used to capture information on patient visits

to hospital and community based ambulatory care: day surgery, outpatient clinics and emergency departments. Client visit data are collected at the time of service in participating facilities and data collection methods may vary by facility. Participation in NACRS did not start simultaneously at all sites or for all types of visit functional centres. Currently, data submission is primarily from the Province of Ontario.

- **OCCC** *(fiscal 1991/92 to 2016/17)*: The Ontario Crohn's and Colitis Cohort Database (OCCC) includes all Ontario patients who were diagnosed with Crohn's disease or Ulcerative Colitis which means Inflammatory Bowel Disease (IBD) when they were aged 0-105 years, diagnosed since 1991. The database was created using hospital discharge abstracts from CIHI (including same day surgery, non-acute care) database, physician service claims from OHIP database, all visits from NACRS database and ODB claims and information regarding the demographics of persons eligible for health care coverage in Ontario from the Registered Persons Database (RPDB). Also, a person has to meet the case definition criteria in order to be included in this database; therefore, they could have been diagnosed many years prior to the data availability. The data have been validated [16], and the sensitivity and specificity for individuals under 18 years of age were 91.1% and 99.5%, respectively.
- **ODD** *(fiscal 1991/92 to 2016/17)*: The Ontario Diabetes Database (ODD) contains all Ontario diabetic patients identified since 1991. The database was created using hospital discharge abstracts from DAD database, Ontario drug benefit claims from ODB database, physician service claims from OHIP database and information regarding the demographics of persons eligible for health care coverage in Ontario from the Registered Persons Database (RPDB). Patients enter the cohort when they

met the case definition criteria; therefore, patients may have DM claim prior to their diagnosis date, but it does not meet the criteria for DM definition. The data have been validated [111], and the sensitivity and specificity for individuals under 18 years of age were 82.8% and 98.9%, respectively.

- **OHIP** *(fiscal 1991/92 to 2016/17)*: OHIP claims data received by ICES contains most claims paid for by the Ontario Health Insurance Plan. The data cover all health care providers who can claim under OHIP (this includes physicians, groups, laboratories, and out-of-province providers). Approximately 95% of specialists and 50% of primary care physicians receive the majority of their income from fee-for-service (FFS). However, all physicians, with the exception of the few hundred family physicians who work in Community Health Centres, are required to submit shadow billings for their non-FFS services. A shadow-billing claim is identical to an FFS claim except that the payment amount is $0.00. Physicians are often provided with cash incentives to encourage them to shadow-bill. Requiring physicians to shadow-bill helps to ensure that the OHIP data accurately (more-or-less) reflect the utilization of physician services in Ontario. Services which are missing from the OHIP claims data include some lab services, services received in provincial psychiatric hospitals, services provided by health service organizations and other alternate funding plans, diagnostic procedures performed on an inpatient basis (e.g. radiology, ECGs), lab services performed at hospitals (both inpatient and same day). For example, a family doctor who belongs to a capitated practice may receive a set sum of money for each patient on the roster, based on the patient's age and sex and does not shadow bill. However, the capitation fee does not cover lab tests. Hence, the OHIP database presumably has a record of all lab tests, and visits to specialists, but may not have a record of all visits

made to the family doctor. Each record represents a single service, identified by the fee code, although that service may be performed several times. For example, if a child has a myringotomy and an adenoidectomy during the surgery, this will generate two records, one for the myringotomy and one for the adenoidectomy.

- **PCCF**: Given a dataset with postal codes, this macro will link to the PCCF files to find other census geographic identifiers such as Dissemination Area/Enumeration Area, Census Division, and Latitude/Longitude; as well as an urban/rural flag and neighborhood income quintile.
- **PM2.5/NO2** *(fiscal 2005/06 to 2016/17)*: The pollution data that contain the level of PM2.5 and NO2 particles in the air.
- **RPDB** *(fiscal 1990/91 to 2016/17)*: The Registered Persons Data Base (RPDB) files provides basic demographic information about anyone who has ever received an Ontario health card number. Data are received directly from the Ministry of Health (MOH) office in Kingston and enriched using in-house ICES core datasets. The files are not perfect. For instance, there are individuals in the "rpdbdemo" file who do not appear in another RPDB data file as individuals are not forced to inform the Ministry when they move and it is likely that the latter data file does not know about people who have moved out of the province (and therefore are no longer eligible for coverage). Hence, the postal code data file for people who move out of the province is not updated and is not really an accurate reflection of where an individual lives.
- **SDS** *(fiscal 2000/01 to 2016/17)*: The Same Day Surgery (SDS) data sets contain patient-level data for day surgery institutions in Ontario. SDS is maintained by the

Canadian Institute for Health Information (CIHI). Note that since April 2003, SDS data are derived from NACRS.

Appendix IV. Variables of Dataset Created for Case Studies

Name	Label	Purpose
ikn	*ICES key number*	(Linkage)
dthdate	*Death date from admin+RPDB data*	(Intermediary)
sex	*Sex in RPDB*	Factor: sex
b_ikn	*Newborn's IKN*	(Linkage)
b_valikn	*Indicator for valid baby IKN*	(Linkage)
m_ikn	*Mom's IKN*	(Linkage)
m_valikn	*Indicator for valid mom IKN*	(Linkage)
m_byear	*Mother's birth year*	(Intermediary)
b_sex	*Baby's sex*	Factor: sex
gest	*Gestational age at birth (weeks)*	Factor: late premature birth
bmi	*Calculated: Mother's BMI*	Factor: high BMI
deltype	*Delivery type*	Factor: c-section
had_influenza	*Evidence of laboratory confirmed H1N1/seasonal influenza and/or an influenza-like illness at any time during the pregnancy*	Effects of prenatal influenza
intbf	*Intention to breastfeed*	Factor: breastfeeding, cow's milk formula (doesn't show mixed feeding, so if the value is TRUE the baby may still have taken formula)
labtype	*Labour type*	Factor: complicated delivery (receiving induction)
mathgtm	*Height in meters*	(Intermediary)
mathp	*Maternal health problem(s)*	Effects of maternal problems
mathp1	*Maternal health problem - Alcohol dependence syndrome/Alcoholism*	Effects of alcohol consumption
mathp2	*Maternal health problem - Asthma*	Factor: maternal asthma, family history
mathp4	*Maternal health problem - Diabetes insulin dependant*	Effects of maternal diabetes insulin
mathp5	*Maternal health problem - Diabetes non-insulin dependant*	Effects of maternal diabetes non-insulin
mathp20	*Psychiatric disorders - Anxiety during this pregnancy*	Factor: anxiety
mathp21	*Psychiatric disorders - Depression during this pregnancy*	Factor: depression
mathp23	*Psychiatric disorders - Previous history of anxiety*	Factor: anxiety
mathp24	*Psychiatric disorders - Previous history of depression*	Factor: depression
mathp26	*Maternal health problem - Thyroid disease*	Factor: thyroid supplements
matwgtkg	*Mother's weight - kg*	Factor: maternal weight gain

multgest	*Multiple gestation*	Effects of multi gestation
nbres	*Newborn Resuscitation*	Effects of reviving newborn
obcomp	*Obstetrical complications*	Factor: complicated delivery
parity	*Parity*	Factor: firstborn child, household crowding, personal towel
phototh	*Phototherapy*	Factor: jaundice
received_antiviral	*Evidence that the mother received antiviral medication at any time during the pregnancy*	Effects of antiviral medications
repass	*Reproductive assistance*	Effects of reproductive assistance
smoking	*Smoking (during preg)*	Factor: maternal smoking, passive smoking exposure
vaccination_influenza	*Evidence that the mother received vaccination for prevention of H1N1 influenza or seasonal influenza at any time during the pregnancy*	Effects of influenza vaccinations
warn	*Warning on linkage of baby's record to mom's*	(Linkage)
b_byear	*Baby's birth year*	(Intermediary)
b_bmonth	*Baby's birth month*	Factor: fungal spore season, grass pollen exposure (Apr to Jul), sunlight exposure
m_bmonth	*Mother's birth month*	Effects of mother's birth month
fsa	*Forward sortation area*	(Intermediary)
acp	*Antenatal care provider*	Effects of having expert care for mother's issues before birth
assisted	*Forceps / vacuum*	Factor: complicated delivery (assisted tools)
augment	*Augmentation*	Factor: complicated delivery (labour augmentation)
csind14	*Indication for C-section - Pre-eclampsia*	(Intermediary)
indind11	*Indication for induction - Pre-eclampsia*	(Intermediary)
obcomp1	*Obstetrical complications - Eclampsia*	(Intermediary)
obcomp9	*Obstetrical complications - Pre-eclampsia*	(Intermediary)
bweight	*Birth weight (grams)*	Factor: low birth weight
PRE_ECLAMPSIA	*Pre-eclampsia*	Factor: pre-eclampsia
B_IBD_INC	*Child's incidence of IBD*	Outcome
B_IBD_DX_LATEST	*Child: Crohn's vs. UC based on latest OHIP*	More specific investigations
B_ASTHMA_INC	*Child's incidence of asthma*	Outcome
B_T1DM_INC	*Child's incidence of T1DM*	Outcome
M_IBD_PRV	*Mother's prevalence of IBD*	Factors related to family history
M_IBD_DX_LATEST	*Mother: Crohn's vs. UC based on latest OHIP*	More specific investigations

M_ASTHMA_PRV	*Mother's prevalence of asthma*	Factor: maternal asthma, family history
M_T1DM_PRV	*Mother's prevalence of T1DM*	Factors related to family history
B_SARD_JIA_INC	*Child's incidence of SARD + JIA*	Outcome
B_MS_INC	*Child's incidence of MS*	Outcome
M_SARD_JIA_PRV	*Mother's prevalence of SARD + JIA*	Factors related to family history
M_MS_PRV	*Mother's prevalence of MS*	Factors related to family history
B_IMMUNE_INC	*Child's incidence of any immune disease*	Outcome
M_IMMUNE_PRV	*Mother's prevalence of any immune disease*	Factors related to family history
srlhinsr_birth	*Beset yearly Sub-region unique identifier (Sub-region LHIN||Sub-region) at birth*	Factor: sunlight exposure/latitude
incquint_birth	*Nearest Census Based Neighbourhood Income Quintile at birth*	Factor: high-income lifestyle
rio2008_birth	*2008 Rurality Index for Ontario at birth*	Factor: living in farm, urbanization
srlhinsr_IBD	*Beset yearly Sub-region unique identifier (Sub-region LHIN||Sub-region) at incidence of IBD*	More specific investigations
incquint_IBD	*Nearest Census Based Neighbourhood Income Quintile at incidence of IBD*	More specific investigations
rio2008_IBD	*2008 Rurality Index for Ontario at incidence of IBD*	More specific investigations
srlhinsr_ASTHMA	*Beset yearly Sub-region unique identifier (Sub-region LHIN||Sub-region) at incidence of asthma*	More specific investigations
incquint_ASTHMA	*Nearest Census Based Neighbourhood Income Quintile at incidence of asthma*	More specific investigations
rio2008_ASTHMA	*2008 Rurality Index for Ontario at incidence of asthma*	More specific investigations
srlhinsr_T1DM	*Beset yearly Sub-region unique identifier (Sub-region LHIN||Sub-region) at incidence of T1DM*	More specific investigations
incquint_T1DM	*Nearest Census Based Neighbourhood Income Quintile at incidence of T1DM*	More specific investigations
rio2008_T1DM	*2008 Rurality Index for Ontario at incidence of T1DM*	More specific investigations
srlhinsr_SARD_JIA	*Beset yearly Sub-region unique identifier (Sub-region LHIN||Sub-region) at incidence of SARD + JIA*	More specific investigations
incquint_SARD_JIA	*Nearest Census Based Neighbourhood Income Quintile at incidence of SARD + JIA*	More specific investigations
rio2008_SARD_JIA	*2008 Rurality Index for Ontario at incidence of SARD + JIA*	More specific investigations
srlhinsr_MS	*Beset yearly Sub-region unique identifier (Sub-region LHIN||Sub-region) at incidence of MS*	More specific investigations

incquint_MS	*Nearest Census Based Neighbourhood Income Quintile at incidence of MS*	More specific investigations
rio2008_MS	*2008 Rurality Index for Ontario at incidence of MS*	More specific investigations
srlhinsr_IMMUNE	*Beset yearly Sub-region unique identifier (Sub-region LHIN||Sub-region) at incidence of first immune disease*	More specific investigations
incquint_IMMUNE	*Nearest Census Based Neighbourhood Income Quintile at incidence of first immune disease*	More specific investigations
rio2008_IMMUNE	*2008 Rurality Index for Ontario at incidence of first immune disease*	More specific investigations
year_concept	*Conception year*	(Intermediary)
srlhinsr_PREG	*Beset yearly Sub-region unique identifier (Sub-region LHIN||Sub-region) at conception*	Factor: sunlight exposure/latitude
incquint_PREG	*Nearest Census Based Neighbourhood Income Quintile at conception*	Factor: high-income lifestyle
rio2008_PREG	*2008 Rurality Index for Ontario at conception*	Factor: living in farm, urbanization
DTH	*Baby died*	More specific investigations
M_age	*Mother's age at birth*	Factor: maternal age
DTH_AGE	*Baby's death age at birth*	More specific investigations
fcob	*Mother's country of birth*	More specific investigations
fsrce	*Mother's source area*	Mother/family's origin
M_IS_IMMIGRANT	*Mother is immigrant*	Factor: 2nd-gen immigrant
pm25_birth	*Air pollution level of PM2.5 in region of residence at birth*	Factor: air pollution, traffic-related particles, fine particulates
pm25_IBD	*Air pollution level of PM2.5 in region of residence at incidence of IBD*	More specific investigations
pm25_asthma	*Air pollution level of PM2.5 in region of residence at incidence of asthma*	More specific investigations
pm25_T1DM	*Air pollution level of PM2.5 in region of residence at incidence of T1DM*	More specific investigations
pm25_SARD_JIA	*Air pollution level of PM2.5 in region of residence at incidence of SARD + JIA*	More specific investigations
pm25_MS	*Air pollution level of PM2.5 in region of residence at incidence of MS*	More specific investigations
pm25_immune	*Air pollution level of PM2.5 in region of residence at incidence of first immune disease*	More specific investigations
pm25_preg	*Air pollution level of PM2.5 in region of residence during pregnancy*	Factor: air pollution, traffic-related particles, fine particulates
no2_birth	*Air pollution level of NO2 in region of residence at birth*	Factor: air pollution, traffic-related particles
no2_IBD	*Air pollution level of NO2 in region of residence at incidence of IBD*	More specific investigations

no2_asthma	*Air pollution level of NO2 in region of residence at incidence of asthma*	More specific investigations
no2_T1DM	*Air pollution level of NO2 in region of residence at incidence of T1DM*	More specific investigations
no2_SARD_JIA	*Air pollution level of NO2 in region of residence at incidence of SARD + JIA*	More specific investigations
no2_MS	*Air pollution level of NO2 in region of residence at incidence of MS*	More specific investigations
no2_immune	*Air pollution level of NO2 in region of residence at incidence of first immune disease*	More specific investigations
no2_preg	*Air pollution level of NO2 in region of residence during pregnancy*	Factor: air pollution, traffic-related particles
M_IBD_AGE_BBIRTH	*Mother's age in days at first diagnosis of IBD counted from baby's birth*	More specific investigations - mother is diagnosed before baby's birth
M_IBD_AGE_MBIRTH	*Mother's age in days at first diagnosis of IBD counted from mother's birth*	More specific investigations - mother's age at diagnosis
M_ASTHMA_AGE_ BBIRTH	*Mother's age in days at first diagnosis of asthma counted from baby's birth*	More specific investigations - mother is diagnosed before baby's birth
M_ASTHMA_AGE_ MBIRTH	*Mother's age in days at first diagnosis of asthma counted from mother's birth*	More specific investigations - mother's age at diagnosis
M_T1DM_AGE_ BBIRTH	*Mother's age in days at first diagnosis of T1DM counted from baby's birth*	More specific investigations - mother is diagnosed before baby's birth
M_T1DM_AGE_ MBIRTH	*Mother's age in days at first diagnosis of T1DM counted from mother's birth*	More specific investigations - mother's age at diagnosis
M_SARD_JIA_AGE_ BBIRTH	*Mother's age in days at first diagnosis of SARD + JIA counted from baby's birth*	More specific investigations - mother is diagnosed before baby's birth
M_SARD_JIA_AGE_ MBIRTH	*Mother's age in days at first diagnosis of SARD + JIA counted from mother's birth*	More specific investigations - mother's age at diagnosis
M_MS_AGE_BBIRTH	*Mother's age in days at first diagnosis of MS counted from baby's birth*	More specific investigations - mother is diagnosed before baby's birth
M_MS_AGE_MBIRTH	*Mother's age in days at first diagnosis of MS counted from mother's birth*	More specific investigations - mother's age at diagnosis
M_IMMUNE_AGE_ BBIRTH	*Mother's age in days at first diagnosis of any immune disease counted from baby's birth*	More specific investigations - mother is diagnosed before baby's birth
M_IMMUNE_AGE_ MBIRTH	*Mother's age in days at first diagnosis of any immune disease counted from mother's birth*	More specific investigations - mother's age at diagnosis
mononucleosis_IBD	*Child diagnosed with mononucleosis before incidence of IBD*	Factor: gamma-herper virus infection
mononucleosis_days	*Child's age in days at first diagnosis of mononucleosis*	Factor: gamma-herper virus infection
mononucleosis_ASTHMA	*Child diagnosed with mononucleosis before incidence of asthma*	Factor: gamma-herper virus infection

mononucleosis_MS	*Child diagnosed with mononucleosis before incidence of MS*	Factor: gamma-herper virus infection
mononucleosis_T1DM	*Child diagnosed with mononucleosis before incidence of T1DM*	Factor: gamma-herper virus infection
mononucleosis_SARD_JIA	*Child diagnosed with mononucleosis before incidence of SARD + JIA*	Factor: gamma-herper virus infection
mononucleosis_Immune	*Child diagnosed with mononucleosis before incidence of first immune disease*	Factor: gamma-herper virus infection
Parvo_B19_IBD	*Child diagnosed with parvovirus B19 before incidence of IBD*	Factor: parvovirus b19
Parvo_B19_days	*Child's age in days at first diagnosis of parvovirus B19*	Factor: parvovirus b19
Parvo_B19_ASTHMA	*Child diagnosed with parvovirus B19 before incidence of asthma*	Factor: parvovirus b19
Parvo_B19_MS	*Child diagnosed with parvovirus B19 before incidence of MS*	Factor: parvovirus b19
Parvo_B19_T1DM	*Child diagnosed with parvovirus B19 before incidence of T1DM*	Factor: parvovirus b19
Parvo_B19_SARD_JIA	*Child diagnosed with parvovirus B19 before incidence of SARD + JIA*	Factor: parvovirus b19
Parvo_B19_Immune	*Child diagnosed with parvovirus B19 before incidence of first immune disease*	Factor: parvovirus b19
Chlamydia_IBD	*Child diagnosed with chlamydia before incidence of IBD*	Factor: chlamydia trachomatis bacteria, chlamydophila pneumoniae bacteria
Chlamydia_days	*Child's age in days at first diagnosis of Chlamydia*	Factor: chlamydia trachomatis bacteria, chlamydophila pneumoniae bacteria
Chlamydia_ASTHMA	*Child diagnosed with chlamydia before incidence of asthma*	Factor: chlamydia trachomatis bacteria, chlamydophila pneumoniae bacteria
Chlamydia_MS	*Child diagnosed with chlamydia before incidence of MS*	Factor: chlamydia trachomatis bacteria, chlamydophila pneumoniae bacteria
Chlamydia_T1DM	*Child diagnosed with chlamydia before incidence of T1DM*	Factor: chlamydia trachomatis bacteria, chlamydophila pneumoniae bacteria
Chlamydia_SARD_JIA	*Child diagnosed with chlamydia before incidence of SARD + JIA*	Factor: chlamydia trachomatis bacteria, chlamydophila pneumoniae bacteria
Chlamydia_Immune	*Child diagnosed with chlamydia before incidence of first immune disease*	Factor: chlamydia trachomatis bacteria, chlamydophila pneumoniae bacteria
Bartonella_IBD	*Child diagnosed with bartonella before incidence of IBD*	Factor: bartonella henselae bacteria

Bartonella_days	*Child's age in days at first diagnosis of Bartonella*	Factor: bartonella henselae bacteria
Bartonella_ASTHMA	*Child diagnosed with bartonella before incidence of asthma*	Factor: bartonella henselae bacteria
Bartonella_MS	*Child diagnosed with bartonella before incidence of MS*	Factor: bartonella henselae bacteria
Bartonella_T1DM	*Child diagnosed with bartonella before incidence of T1DM*	Factor: bartonella henselae bacteria
Bartonella_SARD_JIA	*Child diagnosed with bartonella before incidence of SARD + JIA*	Factor: bartonella henselae bacteria
Bartonella_Immune	*Child diagnosed with bartonella before incidence of first immune disease*	Factor: bartonella henselae bacteria
Mycopl_pneum_IBD	*Child diagnosed with mycoplasma pneumonia before incidence of IBD*	Factor: mycoplasma pneumoniae bacteria
Mycopl_pneum_days	*Child's age in days at first diagnosis of mycoplasma pneumonia*	Factor: mycoplasma pneumoniae bacteria
Mycopl_pneum_ASTHMA	*Child diagnosed with mycoplasma pneumonia before incidence of asthma*	Factor: mycoplasma pneumoniae bacteria
Mycopl_pneum_MS	*Child diagnosed with mycoplasma pneumonia before incidence of MS*	Factor: mycoplasma pneumoniae bacteria
Mycopl_pneum_T1DM	*Child diagnosed with mycoplasma pneumonia before incidence of T1DM*	Factor: mycoplasma pneumoniae bacteria
Mycopl_pneum_SARD_JIA	*Child diagnosed with mycoplasma pneumonia before incidence of SARD + JIA*	Factor: mycoplasma pneumoniae bacteria
Mycopl_pneum_Immune	*Child diagnosed with mycoplasma pneumonia before incidence of first immune disease*	Factor: mycoplasma pneumoniae bacteria
Congen_pneum_IBD	*Child diagnosed with congenital pneumonia before incidence of IBD*	Suggested by domain expert
Congen_pneum_days	*Child's age in days at first diagnosis of congenial pneumonia*	Suggested by domain expert
Congen_pneum_ASTHMA	*Child diagnosed with congenital pneumonia before incidence of asthma*	Suggested by domain expert
Congen_pneum_MS	*Child diagnosed with congenital pneumonia before incidence of MS*	Suggested by domain expert
Congen_pneum_T1DM	*Child diagnosed with congenital pneumonia before incidence of T1DM*	Suggested by domain expert
Congen_pneum_SARD_JIA	*Child diagnosed with congenital pneumonia before incidence of SARD + JIA*	Suggested by domain expert
Congen_pneum_Immune	*Child diagnosed with congenital pneumonia before incidence of first immune disease*	Suggested by domain expert
Congen_Torch_IBD	*Child diagnosed with congenital TORCH before incidence of IBD*	Factor: congenital rubella infection, and suggested by domain expert
Congen_Torch_days	*Child's age in days at first diagnosis of congenital TORCH infection*	Factor: congenital rubella infection, and suggested by domain expert
Congen_Torch_ASTHMA	*Child diagnosed with congenital TORCH before incidence of asthma*	Factor: congenital rubella infection, and suggested by domain expert

Congen_Torch_MS	*Child diagnosed with congenital TORCH before incidence of MS*	Factor: congenital rubella infection, and suggested by domain expert
Congen_Torch_T1DM	*Child diagnosed with congenital TORCH before incidence of T1DM*	Factor: congenital rubella infection, and suggested by domain expert
Congen_Torch_SARD_JIA	*Child diagnosed with congenital TORCH before incidence of SARD + JIA*	Factor: congenital rubella infection, and suggested by domain expert
Congen_Torch_Immune	*Child diagnosed with congenital TORCH before incidence of first immune disease*	Factor: congenital rubella infection, and suggested by domain expert
Echovirus_IBD	*Child diagnosed with echovirus before incidence of IBD*	Factor: echovirus
Echovirus_days	*Child's age in days at first diagnosis of echovirus*	Factor: echovirus
Echovirus_ASTHMA	*Child diagnosed with echovirus before incidence of asthma*	Factor: echovirus
Echovirus_MS	*Child diagnosed with echovirus before incidence of MS*	Factor: echovirus
Echovirus_T1DM	*Child diagnosed with echovirus before incidence of T1DM*	Factor: echovirus
Echovirus_SARD_JIA	*Child diagnosed with echovirus before incidence of SARD + JIA*	Factor: echovirus
Echovirus_Immune	*Child diagnosed with echovirus before incidence of first immune disease*	Factor: echovirus
Cytomegalo_IBD	*Child diagnosed with cytomegalovirus before incidence of IBD*	Factor: cytomegalovirus
Cytomegalo_days	*Child's age in days at first diagnosis of cytomegalovirus*	Factor: cytomegalovirus
Cytomegalo_ASTHMA	*Child diagnosed with cytomegalovirus before incidence of asthma*	Factor: cytomegalovirus
Cytomegalo_MS	*Child diagnosed with cytomegalovirus before incidence of MS*	Factor: cytomegalovirus
Cytomegalo_T1DM	*Child diagnosed with cytomegalovirus before incidence of T1DM*	Factor: cytomegalovirus
Cytomegalo_SARD_JIA	*Child diagnosed with cytomegalovirus before incidence of SARD + JIA*	Factor: cytomegalovirus
Cytomegalo_Immune	*Child diagnosed with cytomegalovirus before incidence of first immune disease*	Factor: cytomegalovirus
Rhinovirus_IBD	*Child diagnosed with rhinovirus before incidence of IBD*	Factor: rhinovirus
Rhinovirus_days	*Child's age in days at first diagnosis of rhinovirus*	Factor: rhinovirus
Rhinovirus_ASTHMA	*Child diagnosed with rhinovirus before incidence of asthma*	Factor: rhinovirus
Rhinovirus_MS	*Child diagnosed with rhinovirus before incidence of MS*	Factor: rhinovirus
Rhinovirus_T1DM	*Child diagnosed with rhinovirus before incidence of T1DM*	Factor: rhinovirus
Rhinovirus_SARD_JIA	*Child diagnosed with rhinovirus before incidence of SARD + JIA*	Factor: rhinovirus

Rhinovirus_Immune	*Child diagnosed with rhinovirus before incidence of first immune disease*	Factor: rhinovirus
Retrovirus_IBD	*Child diagnosed with retrovirus before incidence of IBD*	Factor: retrovirus
Retrovirus_days	*Child's age in days at first diagnosis of retrovirus*	Factor: retrovirus
Retrovirus_ASTHMA	*Child diagnosed with retrovirus before incidence of asthma*	Factor: retrovirus
Retrovirus_MS	*Child diagnosed with retrovirus before incidence of MS*	Factor: retrovirus
Retrovirus_T1DM	*Child diagnosed with retrovirus before incidence of T1DM*	Factor: retrovirus
Retrovirus_SARD_JIA	*Child diagnosed with retrovirus before incidence of SARD + JIA*	Factor: retrovirus
Retrovirus_Immune	*Child diagnosed with retrovirus before incidence of first immune disease*	Factor: retrovirus
Rotavirus_IBD	*Child diagnosed with rotavirus before incidence of IBD*	Factor: rotavirus
Rotavirus_days	*Child's age in days at first diagnosis of rotavirus*	Factor: rotavirus
Rotavirus_ASTHMA	*Child diagnosed with rotavirus before incidence of asthma*	Factor: rotavirus
Rotavirus_MS	*Child diagnosed with rotavirus before incidence of MS*	Factor: rotavirus
Rotavirus_T1DM	*Child diagnosed with rotavirus before incidence of T1DM*	Factor: rotavirus
Rotavirus_SARD_JIA	*Child diagnosed with rotavirus before incidence of SARD + JIA*	Factor: rotavirus
Rotavirus_Immune	*Child diagnosed with rotavirus before incidence of first immune disease*	Factor: rotavirus
Mycobac_Tuber_IBD	*Child diagnosed with mycobacterium tuberculosis before incidence of IBD*	Factor: mycobacterium tuberculosis bacteria
Mycobac_Tuber_days	*Child's age in days at first diagnosis of mycobacterium tuberculosis*	Factor: mycobacterium tuberculosis bacteria
Mycobac_Tuber_ASTHMA	*Child diagnosed with mycobacterium tuberculosis before incidence of asthma*	Factor: mycobacterium tuberculosis bacteria
Mycobac_Tuber_MS	*Child diagnosed with mycobacterium tuberculosis before incidence of MS*	Factor: mycobacterium tuberculosis bacteria
Mycobac_Tuber_T1DM	*Child diagnosed with mycobacterium tuberculosis before incidence of T1DM*	Factor: mycobacterium tuberculosis bacteria
Mycobac_Tuber_SARD_JIA	*Child diagnosed with mycobacterium tuberculosis before incidence of SARD + JIA*	Factor: mycobacterium tuberculosis bacteria
Mycobac_Tuber_Immune	*Child diagnosed with mycobacterium tuberculosis before incidence of first immune disease*	Factor: mycobacterium tuberculosis bacteria
Pseudo_aerug_IBD	*Child diagnosed with pseudomonas aeruginosa before incidence of IBD*	Factor: pseudomonas aeruginosa bacteria
Pseudo_aerug_days	*Child's age in days at first diagnosis of pseudomonas aeruginosa*	Factor: pseudomonas aeruginosa bacteria
Pseudo_aerug_ASTHMA	*Child diagnosed with pseudomonas aeruginosa before incidence of asthma*	Factor: pseudomonas aeruginosa bacteria

Pseudo_aerug_MS	*Child diagnosed with pseudomonas aeruginosa before incidence of MS*	Factor: pseudomonas aeruginosa bacteria
Pseudo_aerug_T1DM	*Child diagnosed with pseudomonas aeruginosa before incidence of T1DM*	Factor: pseudomonas aeruginosa bacteria
Pseudo_aerug_SARD_JIA	*Child diagnosed with pseudomonas aeruginosa before incidence of SARD + JIA*	Factor: pseudomonas aeruginosa bacteria
Pseudo_aerug_Immune	*Child diagnosed with pseudomonas aeruginosa before incidence of first immune disease*	Factor: pseudomonas aeruginosa bacteria
Yersinia_ent_IBD	*Child diagnosed with Yersinia enterocolitica before incidence of IBD*	Factor: yersinia enterocolitica bacteria
Yersinia_ent_days	*Child's age in days at first diagnosis of Yersinia enterocolitica*	Factor: yersinia enterocolitica bacteria
Yersinia_ent_ASTHMA	*Child diagnosed with Yersinia enterocolitica before incidence of asthma*	Factor: yersinia enterocolitica bacteria
Yersinia_ent_MS	*Child diagnosed with Yersinia enterocolitica before incidence of MS*	Factor: yersinia enterocolitica bacteria
Yersinia_ent_T1DM	*Child diagnosed with Yersinia enterocolitica before incidence of T1DM*	Factor: yersinia enterocolitica bacteria
Yersinia_ent_SARD_JIA	*Child diagnosed with Yersinia enterocolitica before incidence of SARD + JIA*	Factor: yersinia enterocolitica bacteria
Yersinia_ent_Immune	*Child diagnosed with Yersinia enterocolitica before incidence of first immune disease*	Factor: yersinia enterocolitica bacteria
Gastro_treat_IBD	*Child had treatment for gastroesophageal reflux disease (GERD) before incidence of IBD*	Factor: gastro-oesophageal reflux treatment (check diagnosis)
Gastro_treat_days	*Child's age in days at first diagnosis of gastroesophageal reflux disease (GERD)*	Factor: gastro-oesophageal reflux treatment (check diagnosis)
Gastro_treat_ASTHMA	*Child had treatment for gastroesophageal reflux disease (GERD) before incidence of asthma*	Factor: gastro-oesophageal reflux treatment (check diagnosis)
Gastro_treat_MS	*Child had treatment for gastroesophageal reflux disease (GERD) before incidence of MS*	Factor: gastro-oesophageal reflux treatment (check diagnosis)
Gastro_treat_T1DM	*Child had treatment for gastroesophageal reflux disease (GERD) before incidence of T1DM*	Factor: gastro-oesophageal reflux treatment (check diagnosis)
Gastro_treat_SARD_JIA	*Child had treatment for gastroesophageal reflux disease (GERD) before incidence of SARD + JIA*	Factor: gastro-oesophageal reflux treatment (check diagnosis)
Gastro_treat_Immune	*Child had treatment for gastroesophageal reflux disease (GERD) before incidence of first immune disease*	Factor: gastro-oesophageal reflux treatment (check diagnosis)
Vir_wheezing_IBD	*Child diagnosed with virus positive wheezing illness before incidence of IBD*	Factor: virus positive wheezing illness (used RSV algorithm)
Vir_wheezing_days	*Child's age in days at first diagnosis of virus positive wheezing illness*	Factor: virus positive wheezing illness (used RSV algorithm)

Vir_wheezing_ASTHMA	*Child diagnosed with virus positive wheezing illness before incidence of asthma*	Factor: virus positive wheezing illness (used RSV algorithm)
Vir_wheezing_MS	*Child diagnosed with virus positive wheezing illness before incidence of MS*	Factor: virus positive wheezing illness (used RSV algorithm)
Vir_wheezing_T1DM	*Child diagnosed with virus positive wheezing illness before incidence of T1DM*	Factor: virus positive wheezing illness (used RSV algorithm)
Vir_wheezing_SARD_JIA	*Child diagnosed with virus positive wheezing illness before incidence of SARD + JIA*	Factor: virus positive wheezing illness (used RSV algorithm)
Vir_wheezing_Immune	*Child diagnosed with virus positive wheezing illness before incidence of first immune disease*	Factor: virus positive wheezing illness (used RSV algorithm)
Opiates_IBD	*Child took opiates before incidence of IBD*	Factor: opiates
Opiates_days	*Child's age in days at first diagnosis of opiates*	Factor: opiates
Opiates_ASTHMA	*Child took opiates before incidence of asthma*	Factor: opiates
Opiates_MS	*Child took opiates before incidence of MS*	Factor: opiates
Opiates_T1DM	*Child took opiates before incidence of T1DM*	Factor: opiates
Opiates_SARD_JIA	*Child took opiates before incidence of SARD + JIA*	Factor: opiates
Opiates_Immune	*Child took opiates before incidence of first immune disease*	Factor: opiates
Gen_Child_Inf_IBD	*Child diagnosed with general childhood infections before incidence of IBD*	Factor: general childhood infection
Gen_Child_Inf_days	*Child's age in days at first diagnosis of general childhood infection*	Factor: general childhood infection
Gen_Child_Inf_ASTHMA	*Child diagnosed with general childhood infections before incidence of asthma*	Factor: general childhood infection
Gen_Child_Inf_MS	*Child diagnosed with general childhood infections before incidence of MS*	Factor: general childhood infection
Gen_Child_Inf_T1DM	*Child diagnosed with general childhood infections before incidence of T1DM*	Factor: general childhood infection
Gen_Child_Inf_SARD_JIA	*Child diagnosed with general childhood infections before incidence of SARD + JIA*	Factor: general childhood infection
Gen_Child_Inf_Immune	*Child diagnosed with general childhood infections before incidence of first immune disease*	Factor: general childhood infection
Antibiotics_IBD	*Child took antibiotics before incidence of IBD*	Factor: use of antibiotics
Antibiotics_days	*Child's age in days at first diagnosis of antibiotics*	Factor: use of antibiotics
Antibiotics_ASTHMA	*Child took antibiotics before incidence of asthma*	Factor: use of antibiotics
Antibiotics_MS	*Child took antibiotics before incidence of MS*	Factor: use of antibiotics

Antibiotics_T1DM	*Child took antibiotics before incidence of T1DM*	Factor: use of antibiotics
Antibiotics_SARD_JIA	*Child took antibiotics before incidence of SARD + JIA*	Factor: use of antibiotics
Antibiotics_Immune	*Child took antibiotics before incidence of first immune disease*	Factor: use of antibiotics
Gastro_inf_IBD	*Child diagnosed with gastroenteritis before incidence of IBD*	Factor: gastroenteritis
Gastro_inf_days	*Child's age in days at first diagnosis of gastroenteritis*	Factor: gastroenteritis
Gastro_inf_ASTHMA	*Child diagnosed with gastroenteritis before incidence of asthma*	Factor: gastroenteritis
Gastro_inf_MS	*Child diagnosed with gastroenteritis before incidence of MS*	Factor: gastroenteritis
Gastro_inf_T1DM	*Child diagnosed with gastroenteritis before incidence of T1DM*	Factor: gastroenteritis
Gastro_inf_SARD_JIA	*Child diagnosed with gastroenteritis before incidence of SARD + JIA*	Factor: gastroenteritis
Gastro_inf_Immune	*Child diagnosed with gastroenteritis before incidence of first immune disease*	Factor: gastroenteritis
Respir_inf_IBD	*Child diagnosed with respiratory infection before incidence of IBD*	Factor: respiratory infection
Respir_inf_days	*Child's age in days at first diagnosis of respiratory infection*	Factor: respiratory infection
Respir_inf_ASTHMA	*Child diagnosed with respiratory infection before incidence of asthma*	Factor: respiratory infection
Respir_inf_MS	*Child diagnosed with respiratory infection before incidence of MS*	Factor: respiratory infection
Respir_inf_T1DM	*Child diagnosed with respiratory infection before incidence of T1DM*	Factor: respiratory infection
Respir_inf_SARD_JIA	*Child diagnosed with respiratory infection before incidence of SARD + JIA*	Factor: respiratory infection
Respir_inf_Immune	*Child diagnosed with respiratory infection before incidence of first immune disease*	Factor: respiratory infection
Rhinitis_IBD	*Child diagnosed with allergic rhinitis before incidence of IBD*	Factor: rhinitis
Rhinitis_days	*Child's age in days at first diagnosis of Rhinitis*	Factor: rhinitis
Rhinitis_ASTHMA	*Child diagnosed with allergic rhinitis before incidence of asthma*	Factor: rhinitis
Rhinitis_MS	*Child diagnosed with allergic rhinitis before incidence of MS*	Factor: rhinitis
Rhinitis_T1DM	*Child diagnosed with allergic rhinitis before incidence of T1DM*	Factor: rhinitis
Rhinitis_SARD_JIA	*Child diagnosed with allergic rhinitis before incidence of SARD + JIA*	Factor: rhinitis
Rhinitis_Immune	*Child diagnosed with allergic rhinitis before incidence of first immune disease*	Factor: rhinitis
Salmonella_IBD	*Child diagnosed with salmonella before incidence of IBD*	Factor: salmonella
Salmonella_days	*Child's age in days at first diagnosis of Salmonella*	Factor: salmonella

Salmonella_ASTHMA	*Child diagnosed with salmonella before incidence of asthma*	Factor: salmonella
Salmonella_MS	*Child diagnosed with salmonella before incidence of MS*	Factor: salmonella
Salmonella_T1DM	*Child diagnosed with salmonella before incidence of T1DM*	Factor: salmonella
Salmonella_SARD_JIA	*Child diagnosed with salmonella before incidence of SARD + JIA*	Factor: salmonella
Salmonella_Immune	*Child diagnosed with salmonella before incidence of first immune disease*	Factor: salmonella
Pertussis_IBD	*Child diagnosed with pertussis before incidence of IBD*	Factor: pertussis
Pertussis_days	*Child's age in days at first diagnosis of Pertussis*	Factor: pertussis
Pertussis_ASTHMA	*Child diagnosed with pertussis before incidence of asthma*	Factor: pertussis
Pertussis_MS	*Child diagnosed with pertussis before incidence of MS*	Factor: pertussis
Pertussis_T1DM	*Child diagnosed with pertussis before incidence of T1DM*	Factor: pertussis
Pertussis_SARD_JIA	*Child diagnosed with pertussis before incidence of SARD + JIA*	Factor: pertussis
Pertussis_Immune	*Child diagnosed with pertussis before incidence of first immune disease*	Factor: pertussis
Strep_pyo_IBD	*Child diagnosed with Streptococcus pyogenes before incidence of IBD*	Factor: streptococcus pyogenes bacteria
Strep_pyo_days	*Child's age in days at first diagnosis of Streptococcus pyogenes*	Factor: streptococcus pyogenes bacteria
Strep_pyo_ASTHMA	*Child diagnosed with Streptococcus pyogenes before incidence of asthma*	Factor: streptococcus pyogenes bacteria
Strep_pyo_MS	*Child diagnosed with Streptococcus pyogenes before incidence of MS*	Factor: streptococcus pyogenes bacteria
Strep_pyo_T1DM	*Child diagnosed with Streptococcus pyogenes before incidence of T1DM*	Factor: streptococcus pyogenes bacteria
Strep_pyo_SARD_JIA	*Child diagnosed with Streptococcus pyogenes before incidence of SARD + JIA*	Factor: streptococcus pyogenes bacteria
Strep_pyo_Immune	*Child diagnosed with Streptococcus pyogenes before incidence of first immune disease*	Factor: streptococcus pyogenes bacteria
Coxsackie_B_IBD	*Child diagnosed with coxsackie B virus before incidence of IBD*	Factor: coxsackie b virus
Coxsackie_B_days	*Child's age in days at first diagnosis of coxsackie B virus*	Factor: coxsackie b virus
Coxsackie_B_ASTHMA	*Child diagnosed with coxsackie B virus before incidence of asthma*	Factor: coxsackie b virus
Coxsackie_B_MS	*Child diagnosed with coxsackie B virus before incidence of MS*	Factor: coxsackie b virus
Coxsackie_B_T1DM	*Child diagnosed with coxsackie B virus before incidence of T1DM*	Factor: coxsackie b virus
Coxsackie_B_SARD_JIA	*Child diagnosed with coxsackie B virus before incidence of SARD + JIA*	Factor: coxsackie b virus

Coxsackie_B_Immune	*Child diagnosed with coxsackie B virus before incidence of first immune disease*	Factor: coxsackie b virus
Mumps_IBD	*Child diagnosed with mumps before incidence of IBD*	Factor: mumps
Mumps_days	*Child's age in days at first diagnosis of mumps*	Factor: mumps
Mumps_ASTHMA	*Child diagnosed with mumps before incidence of asthma*	Factor: mumps
Mumps_MS	*Child diagnosed with mumps before incidence of MS*	Factor: mumps
Mumps_T1DM	*Child diagnosed with mumps before incidence of T1DM*	Factor: mumps
Mumps_SARD_JIA	*Child diagnosed with mumps before incidence of SARD + JIA*	Factor: mumps
Mumps_Immune	*Child diagnosed with mumps before incidence of first immune disease*	Factor: mumps
Thyroid_supp_IBD	*Child took thyroid supplements before incidence of IBD*	Factor: thyroid supplements (check diagnosis)
Thyroid_supp_days	*Child's age in days when first took thyroid supplements*	Factor: thyroid supplements (check diagnosis)
Thyroid_supp_ASTHMA	*Child took thyroid supplements before incidence of asthma*	Factor: thyroid supplements (check diagnosis)
Thyroid_supp_MS	*Child took thyroid supplements before incidence of MS*	Factor: thyroid supplements (check diagnosis)
Thyroid_supp_T1DM	*Child took thyroid supplements before incidence of T1DM*	Factor: thyroid supplements (check diagnosis)
Thyroid_supp_SARD_JIA	*Child took thyroid supplements before incidence of SARD + JIA*	Factor: thyroid supplements (check diagnosis)
Thyroid_supp_Immune	*Child took thyroid supplements before incidence of first immune disease*	Factor: thyroid supplements (check diagnosis)
Atopy_IBD	*Child diagnosed with atopy before incidence of IBD*	Factor: atopy
Atopy_days	*Child's age in days at first diagnosis of atopy*	Factor: atopy
Atopy_ASTHMA	*Child diagnosed with atopy before incidence of asthma*	Factor: atopy
Atopy_MS	*Child diagnosed with atopy before incidence of MS*	Factor: atopy
Atopy_T1DM	*Child diagnosed with atopy before incidence of T1DM*	Factor: atopy
Atopy_SARD_JIA	*Child diagnosed with atopy before incidence of SARD + JIA*	Factor: atopy
Atopy_Immune	*Child diagnosed with atopy before incidence of first immune disease*	Factor: atopy
Appendectomy_IBD	*Child had appendectomy before incidence of IBD*	Factor: appendectomy (diagnosis and treatment)
Appendectomy_days	*Child's age in days at first diagnosis of Appendectomy*	Factor: appendectomy (diagnosis and treatment)
Appendectomy_ASTHMA	*Child had appendectomy before incidence of asthma*	Factor: appendectomy (diagnosis and treatment)
Appendectomy_MS	*Child had appendectomy before incidence of MS*	Factor: appendectomy (diagnosis and treatment)

Appendectomy_T1DM	*Child had appendectomy before incidence of T1DM*	Factor: appendectomy (diagnosis and treatment)
Appendectomy_SARD_JIA	*Child had appendectomy before incidence of SARD + JIA*	Factor: appendectomy (diagnosis and treatment)
Appendectomy_Immune	*Child had appendectomy before incidence of first immune disease*	Factor: appendectomy (diagnosis and treatment)
amniocentesis	*Mother had amniocentesis during pregnancy*	Factor: amniocentesis
b_days_asthma	*Age of child in days at incidence of asthma*	More specific investigations
b_days_ibd	*Age of child in days at incidence of IBD*	More specific investigations
b_days_immune	*Age of child in days at incidence of first immune disease*	More specific investigations
b_days_ms	*Age of child in days at incidence of MS*	More specific investigations
b_days_SARD_JIA	*Age of child in days at incidence of SARD + JIA*	More specific investigations
b_days_T1DM	*Age of child in days at incidence of T1DM*	More specific investigations
m_days_asthma	*Age of mother counted from baby's birth in days at incidence of asthma*	More specific investigations
m_days_ibd	*Age of mother counted from baby's birth in days at incidence of IBD*	More specific investigations
m_days_immune	*Age of mother counted from baby's birth in days at incidence of first immune disease*	More specific investigations
m_immune_date	*Mother's incident date to first immune disease*	(Intermediary)
m_days_ms	*Age of mother counted from baby's birth in days at incidence of MS*	More specific investigations
m_days_SARD_JIA	*Age of mother counted from baby's birth in days at incidence of SARD + JIA*	More specific investigations
m_days_T1DM	*Age of mother counted from baby's birth in days at incidence of T1DM*	More specific investigations

Appendix V. Preprocessing Details of the First Case Study

The preprocessing steps taken to prepare the data for the first case study are presented in this appendix.

Technical Preprocessing

The Technical Preprocessing Guideline was followed. The details of each step are provided in continue.

Step 1. Dataset Creation

BORN-Niday dataset was our base dataset as it already was in a two-dimensional format in which each record represented a newborn in Ontario and the columns featured early life attributes of the baby in addition to mother's neonatal attributes.

Our inclusion criteria were:

- Ontarian newborns (from fiscal 2006 to 2011) and their mothers – been in Ontario 10 months before birth to have OHIP

Our exclusion criteria were:

- Not continuously eligible in OHIP from birth to March 31, 2017, or death, whichever comes first
- Infants who died in the first 7 days of life
- Invalid records, and records inconsistent with reliable population datasets

There were 813,719 records included from the Niday dataset. Table Apx-V.1 demonstrates how 88,089 records were eliminated as a matter of applying the exclusion criteria step by step. Eventually, the study population contained 725,630 eligible infants born in Ontario from April 2006 to March 2012.

Table Apx-V.1: Niday records elimination due to the exclusion criteria

N	Exclusion	Explanation
813,719		
	19,724	IKN is not valid
793,995		
	2,068	Warning on the linkage
791,927		
	411	Doubles with different entries for variables. (Double Records with all variables same are kept in (109))
791,516		
	19,238	Baby birthday or sex is different than RPDB
772,278		
	1,505	Died in 7 days
770,773		
	45,143	Not eligible until March 2017
725,630		

Step 2. Feature Construction

We constructed 294 new features that each represented one or more factors. Most of these features were related to 31 health conditions that contributed as either risk or protected factor on at least one of the outcomes (IMIDs) in the literature and they were identified using the captured diagnostic and procedural codes for each patient in the administrative data. Each of these features had to be constructed separated for every outcome as the index date was different – i.e., if a health condition had occurred after the incident date of an outcome, it means the patient was not exposed to this health condition beforehand and we consider it as

not being diagnosed with this condition, and the incident date is different for every outcome. For patients without any IMID (controls), we considered the recent date in the data (March 31, 2017) as the index date as they do not have any incident date. In addition, if descriptive models were going to be built, there is no outcome (target variable) in this type of modelling to consider its incident date as the index date. Therefore, we constructed an additional version of the feature to explore from birth to recent date as the look-up window and capture the age they were diagnosed with that condition for all patients – the same as what we did for controls in the other versions of the feature. In other words, we considered the lookup window as following to construct comorbidity variables based on the different types of modelling:

- **Descriptive modelling:** there is no outcome, so the exposure is looked up until the latest time in the data to identify whether the individual was exposed at all or not.
- **Predictive modelling:** the exposure is looked up until the point that outcome occurs so identify whether the individual was exposed before the outcome happened or not. For the controls, the outcome did not occur until the latest time in the data, so the exposure is looked up until the latest time in the data to identify whether the individual was exposed at all or not.

Out of all the constructed features for various outcomes, 58 features were specifically prepared for this case study. Others were constructed to act as intermediary in building the features specific to this case study, or were prepared to be available for next investigations and potential case studies. In continue, we provide more information on how each set of these features were constructed using the selected datasets and attributes. Note that we also did include suspect diagnoses in constructing outcome-related and covariate features.

Related to outcome:

- Identified the first diagnosis date of each outcome (IMIDs) for child and mom using the algorithms described in section 6.1.2.3 where the cohorts were defined. The first diagnosis date for the child was the incident date, while the first date found in the data for the mom was the prevalent date as she might had been diagnosed earlier too, but the information was not available in our data. In addition to the features showing child's incident date and mom's prevalent date to each outcome, a binary feature was also constructed for child and mom separately demonstrating whether they were diagnosed with the outcome or not. This was done by simply checking whether there was a diagnosis date for the child and mom or not in each outcome. The union of these binary features for child and mom separately, in addition to the earliest incident and prevent dates identified for the child and mom respectively among all outcomes was considered as the status flag and diagnosis date for the version of features that considered all IMIDs as the outcome (this case study). Child's incidence status to the outcome, mother's prevalence status to the outcome, child's diagnosis age, and mother's diagnosis age (counted from mother's birth time and baby's birth time) were the 5 new features.

Covariates:

- **Mother is immigrant:** Linked the mother's key number to the CIC dataset to identify whether the mother was an immigrant or not. Constructed 1 new feature.
- **Baby's birth month:** Derived from Niday's baby's birth date variable (done by an ICES analyst who had access to level 1 information). Constructed 1 new feature.
- **Mother's birth month:** Derived from Niday's mother's birth date variable (done by an ICES analyst who had access to level 1 information). Constructed 1 new feature.

- **Mother's age at delivery:** Derived from Niday's mother's birth year variable. Constructed 1 new feature.
- **Baby's death age:** Derived from RPDB's death date variable. Constructed 1 new feature.
- **Amniocentesis:** Linked the mother's key number to the OHIP dataset using CCI (Canadian Classification of Health Interventions) procedural/fee codes of J149 and Z778 to identify whether the mother had any amniocentesis during the one year before the birth of baby or not. Constructed 1 new feature.
- **Pre-eclampsia:** Union of CSIND14, INDIND11, OBCOMP1, and OBCOMP9 variables in the Niday dataset. Constructed 1 new feature.
- **Income quintile, rurality index, and sub-region:** Fetched the postal codes from RPDB (done by an ICES analyst who had access to level 1 information) at conception date (birth date minus gestational age weeks), birth date, and incident dates for all concerned outcomes. Then linked to another RPDB dataset to identify the nearest census-based neighborhood income quintile, the 2008 rurality index for Ontario, and the best yearly sub-region unique identifier at conception, at birth time, and at incident times. Constructed 9 new features for this case study.
- **Air quality:** Fetched the postal codes from RPDB the same as described above at conception date, birth date, and outcome incident dates. Then linked to PM2.5 and NO2 datasets (dduque's files in 901.072.000 at ICES) to identify the level of pollution (these two particles) in air during pregnancy (calculated the average of values from conception month to birth month), at birth time, and at incident times. Note that the air

quality raw data contained remarkable amounts of missing values. Constructed 6 new features for this case study.

- **Health conditions:** Linked the child's key number to the DAD and OHIP datasets to find their first record with the relevant diagnosis codes and identify the date that the child was diagnosed with the health conditions listed below. Prepared different versions of each feature (31 of the new features were for this case study) by comparing the extracted diagnosis date with each outcome date to verify if it occurred before the outcome and then entered the child's age at diagnosis time using the following categories:
    - **n/a:** never been diagnosed with the concerned condition (or not diagnosed before the outcome's incident date, in outcome-specific versions).
    - **nb:** diagnosed with the concerned condition for the first time between birth and 1 month old.
    - **1-3m:** diagnosed with the concerned condition for the first time between 1 and 3 months old.
    - **3-12m:** diagnosed with the concerned condition for the first time between 3 and 12 months old.
    - **1:** diagnosed with the concerned condition for the first time at the age of 1 year old.
    - **2:** diagnosed with the concerned condition for the first time at the age of 2 years old.
    - …
- … for these health conditions using the defined diagnosis codes:

- **Infectious mononucleosis:** DAD (DX10CODE) with ICD10: B27.*
- **Parvovirus B19:** DAD (DX10CODE) with ICD10: B34.3, B97.6
- **Chlamydia:** DAD (DX10CODE) with ICD10: A55, A56.*, A70, A71.*, A74.*, J16.0, O98.32, O98.33, O98.319, P23.1, P39.1
- **Bartonella henselae bacteria:** DAD (DX10CODE) with ICD10: A44.*
- **Mycoplasma pneumoniae bacteria:** DAD (DX10CODE) with ICD10: A49.3, B96.0, J15.7, J20.0
- **Congenital pneumonia:** DAD (DX10CODE) with ICD10: P23.6
- **Congenital TORCH infection:** DAD (DX10CODE) with ICD10: A50.*, B01.* (< 3 months old), B06.* (< 3 months old), B34.3, B97.6, P35.0, P35.1, P35.2, P35.3, P35.8, P35.9, P37.1, P39.8
    - TORCH stands for Toxoplasmosis, Other (syphilis, varicella-zoster, parvovirus B19), Rubella, Cytomegalovirus (CMV), and Herpes infections. These are the most common congenital infections. We also included congenital hepatitis and unspecified congenital in the "Other" section to cover all common congenital infections.
- **Echovirus:** DAD (DX10CODE) with ICD10: A08.39, A87.0, B34.1, B97.12, J20.7
- **Cytomegalovirus:** DAD (DX10CODE) with ICD10: B25.*, B27.1*, P35.1
- **Rhinovirus:** DAD (DX10CODE) with ICD10: B34.8, B97.89, J20.6
- **Retrovirus:** DAD (DX10CODE) with ICD10: B33.3, B97.3*
- **Rotavirus:** DAD (DX10CODE) with ICD10: A08.0
- **Mycobacterium tuberculosis:** DAD (DX10CODE) with ICD10: A15.*, A17.*, A18.*, A31.*, O98.0*

- **Pseudomonas aeruginosa:** DAD (DX10CODE) with ICD10: A41.52, B96.5, J15.1, P23.5
- **Yersinia enterocolitica:** DAD (DX10CODE) with ICD10: A04.6, A28.2
- **Gastroesophageal reflux disease (GERD) treatment:** DAD (DX10CODE) with ICD10: K21, K21.0, K21.9
- **Virus positive wheezing illness:** DAD (DX10CODE) with ICD10: B97.4, J12.1, J20.5, J21.0
- **Opiates:** DAD (DX10CODE) with ICD10: F11.*, P96.1, T40.0, T40.1, T40.2, T40.3, T40.4
- **Gastroenteritis:** OHIP (DXCODE) with code: 009
- **Respiratory infection:** OHIP (DXCODE) with code: 460
- **Allergic rhinitis:** OHIP (DXCODE) with code: 477
- **Salmonella:** OHIP (DXCODE) with code: 003
- **Pertussis:** OHIP (DXCODE) with code: 033
- **Streptococcus pyogenes:** OHIP (DXCODE) with code: 034
- **Coxsackie B virus:** OHIP (DXCODE) with code: 074
- **Mumps:** OHIP (DXCODE) with code: 072
- **Thyroid supplements:** OHIP (DXCODE) with code: 243, 244
- **Atopy:** OHIP (DXCODE) with code: 477; or is available in the ASTHMA database
- **General childhood infection:** DAD (DX10CODE) with ICD10: A01.03, A02.22, A15.0, A20.2, A22.1, A31.0, A37.01, A37.11, A37.81, A37.91, A40.3, A49.8, A50.04, A54.84, B01.2, B05.2, B06.81, B25.0, B38.0, B38.1, B39.2, B44.9, B45.0, B49, B59, B77.81, B95.3, B96.0, B96.1, B97.4, B97.89,

G00.1, J10.0*, J10.08, J11.0*, J12.*, J13, J14, J15.*, J16.*, J17, J18.*, J20.0, J20.5, J21.0, J84.11*, J84.2, J84.89, J84.9, J85.1, J95.851, J95.89, M00.1*, P23.*, J09.X1; or OHIP (DXCODE) with code: 052, 381, 382, 460, 463, 464, 466, 486, 590, 595, 597, 599, 601, 968

- We considered otitis media, tonsillitis, upper respiratory tract infections (URTI), lower respiratory tract infections (LRTI), urinary tract, croup, chickenpox, pneumonia, bronchiolitis, and respiratory syncytial virus (RSV) as general childhood infections.
- Used this algorithm [126] to identify RSV.
- Included ICD10 codes with "pneumonia" in their title or synonym terms, and ICD10 codes that are due to pneumonia. Considered pneumonia, pneumoniae, and pneumococcal in keywords, but not pneumoconiosis, pneumonitis, and pneumothorax.

- **Use of antibiotics:** DAD (DX10CODE) with ICD10: A01.03, A02.22, A15.0, A20.2, A22.1, A31.0, A37.01, A37.11, A37.81, A37.91, A40.3, A49.8, A50.04, A54.84, B01.2, B05.2, B06.81, B25.0, B38.0, B38.1, B39.2, B44.9, B45.0, B49, B59, B77.81, B95.3, B96.0, B96.1, B97.89, G00.1, J10.0*, J10.08, J11.0*, J12.*, J13, J14, J15.*, J16.*, J17, J18.*, J20.0, J84.11*, J84.2, J84.89, J84.9, J85.1, J95.851, J95.89, M00.1*, P23.*, J09.X1; or OHIP (DXCODE) with code: 381, 382, 463, 486, 590, 595, 597, 599, 601, 682, 968

  - We considered otitis media, tonsillitis, urinary tract, pneumonia, and cellulitis as diagnoses that lead to the use of antibiotics.
  - Included the same codes for pneumonia as described above under the general childhood infections.

- **Appendectomy surgery:** DAD, SDS (INCODE) with CCI: 1NV89*; or OHIP (FEECODE) with code: S205
    - Used OHIP billing codes for appendectomy and CCI codes for excision total, not drainage of appendix. Sometimes appendicitis doesn't lead to appendectomy. In very young children (<1 year old) an appendectomy might be performed as part of bowel resection for reasons that are not appendicitis. Therefore, we used the procedure codes.

The diagnoses captured in DAD database were marked to be questionable or not. For example, the patient might had been diagnosed with pneumonia, but it was later analyzed that due to other observations, this diagnosis actually might not be pneumonia, so they flagged it as questionable. We excluded these diagnoses when constructing outcome-related (applies to MS and JIA) and health conditions (e.g., general childhood infections) variables that fetched data from DAD data files.

Step 3. Exploration and Planning

We captured all the pre-existing and constructed variables in the created dataset along with the description and purpose of each (i.e., it is related to which factor or reason) in a spreadsheet. Then, we fetched the frequency of all distinct values in every variable and explored the data in each variable to identify the number and percentage of missing values, the number of distinct values, the percentage of values with highest and lowest frequencies, and inconsistent values. Also, we determined the data type (continuous, categorical, or date), data subtype (ordinal, nominal, binary, natural, or real), final dataset (for descriptive modelling or predictive modelling with each outcome being the target variable), and variable type in each dataset (input, target, or auxiliary). Based on the computed statistics for each variable, we were

able to plan the required technical preprocessing actions. List of all 354 variables included in the created dataset with the purpose of each can be found in Appendix IV (Variables of Dataset Created for Case Studies). Note that not all of the 354 variables were selected or constructed for this case study as different versions of some features had to be constructed for various outcomes which were any IMID, asthma, IBD, MS, SARDs, T1DM, and no target for descriptive modelling.

Table Apx-V.2 demonstrated the summary results of data exploration and preprocessing planning. Note that the count of variables in various states could have overlap. For instance, a variable is counted both as duplicate and with high missingness.

Table Apx-V.2: Exploration and planning of created dataset in the first case study

State/Action	*Variable Count*
Total in created dataset	354
Remove as intermediary or linkage	17
Remove as duplicate	9
With high missingness *(frequency >50%)*	95
All or most missing values have valid meaning	78
Remove due to high missingness without valid meaning	20
Remove as unary *(after excluding missing values)*	0
Almost unary *(>95% of values are unique after excluding missing values)*	172
Require imputation *(input variables with missingness without valid meaning)*	40
Might require binning *(input or target variables with more than 100 distinct values)*	32
With noise or inconsistency	9

We derived final datasets from this created dataset that is described above. For the first case study, we prepared the IMM dataset which included 725,630 records and 95 variables. Note that we did not include variables specifically constructed for other outcomes and also, we excluded the intermediary and linkage variables. At the point that the IMM dataset was created, the data types of variables were set based on how we defined in the planning sheet.

Table Apx-V.3: Statistics of the IMM dataset after preprocessing planning

Title	*Count*
Total records	725,630
Input variables	81
Target variables	1
Auxiliary variables	13
Eliminated intermediary and linkage variables	17
Eliminated unrelated variables	242
Extracted CSV file size	*~170 MB*

Step 4. Data Partitioning

As the IMM dataset is large, we used a 70/30 data split partitioning which 70% of it formed the training set and 30% the test set. The size of each set was as following:

- IMM Training Set: 507,941 records
- IMM Test Set: 217,689 records

Step 5. Data Cleaning

The birth weight variable (BWEIGHT) contained strange values which was due to using inconsistent measurement unit. Vast majority of values were in grams; however, more than

2000 records in the IMM dataset had birth weight of less than 1000 grams. After further investigations, we noticed they are not noises and can be fixed. The following criteria was set and applied to the applicable records:

- 1-digit, 2-digit, and 3-digit values (up to 400) were most likely meant to be in kilograms, as their gestational ages were around 40 weeks; so, their number formats were fixed and transformed into 4-digit values.
- 3-digit values from 400 to 1000 were correctly in grams, as their gestational ages were mainly between 22 and 34 weeks which made sense; so, we kept them as is. Note that even though it was mentioned in the source database's information that only birth weight of 500 grams and greater were included, the birth weights between 400 and 500 grams looked valid since their gestational ages were low. It did not make sense for a baby born with 25 weeks of pregnancy to have a birth weight of over 4 kilograms.

Overall, BWEIGHT values in 36 records (25 in training; 11 in test) were fixed:

- 16 values <10
- 7 values <100
- 13 values <400

In addition, value 3 in the intention for breastfeeding variable (INTBF) referred to missing information. Therefore, we changed it to null value to be considered as missing. Overall, INTBF values in 3,079 records (2,155 in training; 924 in test) were changed from 3 to null value.

Step 6. Missing Values

The following actions were done to handle missing values in the listed order based on the directions provided in the guideline.

Input Variables

- **Replace:** The source country of immigration variable (FSRCE) had null values, but they had a valid meaning which was the mother was not an immigrant (or basically her source country is Canada). These null values were generated as a matter of linking the mother's IKN to the immigration dataset (CIC) because the mother was not available on that dataset – i.e., did not have any immigration records which means she was not an immigrant. Therefore, we transformed the null values to 0 to have a valid value. Overall, we replaced FSRCE values in 526,601 records (368,620 in training; 157,981 in test).
- **Remove variable:** We removed the following variables due to high amount of missingness (>50% of values in each variable) and the percentage of missing values in each variable is indicated below.
    - BMI: 85%
    - HAD_INFLUENZA: 83%
    - MATWGTKG: 78%
    - RECEIVED_ANTIVIRAL: 83%
    - VACCINCATION_INFLUENZA: 84%
- **Truncate record:** We truncated 3,587 records (2,510 in training; 1,077 in test) which was equal to 0.5% of total records as they contained high amount of missingness (>25% of input variables or number of input cells in record).
- ***Imputation:*** After removing records and columns with high amount of missingness, in overall 592,130 records (414,491 in testing; 177,639 in test) contained at least one missing cell among the input variables which is around 82% of IMM dataset records.

However, only 3.6% of input variable cells had missing data. As we planned to use k-nearest neighbors (KNN) algorithm in R for imputation, we had to postpone imputation specifically to when categorical variables were one-hot encoded, unqualified variables were removed, and the data were normalized – after "Step 9. Data Normalization". A KNN model was built on the training set to impute missing values, and then the same model was used to impute missing data in the test set. Note that KNN imputation had an exponential computational time as we experimented on various data sizes and the results shown in Figure Apx-V.1. After imputation was done, we ran verification by comparing the data before and after imputation in the training set:

- As the data were standardized (z-score normalization), we expected the number of null values in input variables become 0, and the mean and standard deviation of all input variables still remained (or very close to) 0 and 1, respectively.
- We confirmed that there were no missing values in input variables after imputation.
- However, the mean and standard deviation of most imputed variables had slightly changed. In three variables (AUGMENT, NO2_BIRTH, and NO2_PREG), even though the mean had remained close to 0, the standard deviation was reduced with more than 0.1 unit. The reason was possibly because these three variables had higher amounts of missingness (respectively 192K, 191K, and 180K cells in the IMM dataset). As their standard deviation still had not changed significantly, we preserved them. Note that the amount of missingness in these three variables were still below 50%; therefore, they were not removed earlier.

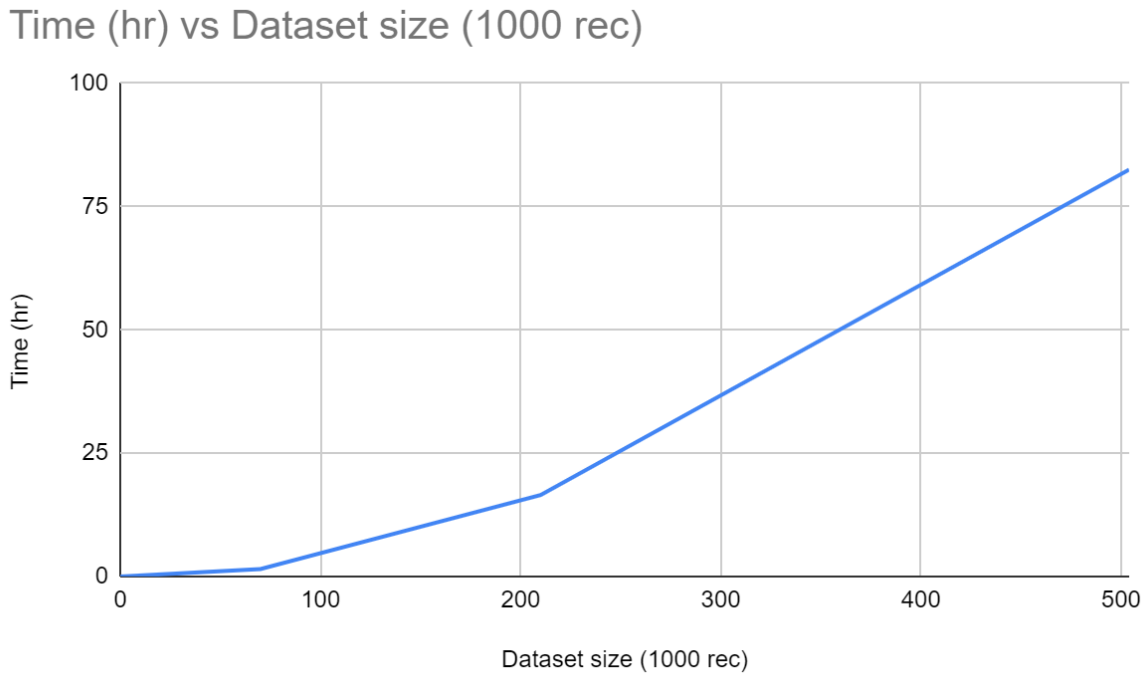

Figure Apx-V.1: Computational time of KNN imputation on different data sizes

Other Variables

There were also some other variables that their missing values had valid meanings – as identified during data exploration and marked in the preprocessing planning sheet –, but none of them were part of the input or target variables in the IMM dataset and did not need any preprocessing action. For instance, the location-based variables – including air quality (NO2 and PM2.5), income quintile, rurality, and sub-region – at the diagnosis time also had high missingness. However, vast majority of this missingness was since the individuals were not diagnosed with the outcome disease, so did not have any diagnosis date to check their location in that date, and subsequently, identify their postal code and location-based information. In other words, the majority of missing values in these variables expressed that these records were part of the controls, and the variables were not applicable to them. The mother's age at

time of child's diagnosis variables had the same issue. So, as their values were only applicable to cases, we preserved them as auxiliary variables for further investigations in the results. Note that they could not be an appropriate input variable in the unseen data too because they were dependent to whether the individual is diagnosed with the outcome or not, while this is the information we were trying to predict and do not know about it in new unseen data. This was very different with the comorbidity variables that we checked the history of patients up to their diagnosis date to verify if they were diagnosed with that comorbidity or not. Even for the controls, we checked the history up to the present (or the latest time in the data). So even if most of these comorbidity variables were almost-unary expressing that most patients were not diagnosed with it, it was a true value and had a meaning within the context of the variable – e.g., the patient did not have mumps.

Step 7. Categorical Variables

All categorical variables were one-hot encoded to binary dummy variables. As a result, the number of input variables were increased from 76 to 587.

Step 8. Feature Reduction

Applied the following actions to remove the unqualified variables:

- **Unary or almost-unary variables:** We removed 500 variables as they were unary or mostly almost-unary in the training set. These variables were removed from the test set later, no matter if they were unary or almost-unary in the test set.
- **Unique variables:** Linkage and intermediary variables were already excluded at the point the IMM dataset was derived from the larger dataset we created in "Step 3. Exploration and Planning". Child and mom unique IDs were part of these excluded variables. There were no other unique variables in the dataset.

- **Highly correlated variables:** We removed 4 variables as they were duplicated or highly correlated to some other input variables in the training set using the Pearson method. Note that imputation of remained missing values had not been done at this point; therefore, we used pairwise complete observation which only included the pair values that did not have missing data to compute correlation. If two variables had a high correlation with each other, the function checked at the mean absolute correlation of each variable and removed the variable with the largest mean absolute correlation value. The same variables were removed from the test set later, no matter if they were correlated in the test set. The variables with duplication or high correlation and their correlation values were as following:
    - B_SEX was a duplicate of SEX (both representing the biological sex of baby)
    - FSRCE.0 had correlation of -0.9999 with M_IS_IMMIGRANT
    - RIO2008_BIRTH had correlation of 0.9243 with RIO2008_PREG
    - ATOPY_IMMUNE.NA had correlation of 1 with RHINITIS_IMMUNE.NA

Step 9. Data Normalization

All input variables in the training set were standardized (z-score normalization) – aka scaled and centered. Eventually, input variables in the test set were also standardized based on the calculations from the training set. Note that imputation of remained missing values occurred at this point which is described in step 6.

Step 10. Data Balancing

As the IMM training set was large, we decided to use undersampling technique to balance it. Table Apx-V.4 shows the number of variables and records in the IMM training set that was used to build predictive models for the first case study.

Table Apx-V.4: Statistics of balancing IMM training set

Title	*Count*
Total variables	97
Input variables	83
Total records	505,431
Before balancing	class-1: 85,949 / class-0: 419,482
After balancing	class-1: 85,949 / class-0: 85,949
Final total records	171,898

Step 11. Resampling Data

As we were dealing with large number of data points and computationally expensive classifiers such as neural networks, we decided to use the 10-fold cross validation (CV) resampling in building predictive models.

Step 12. Runtime Filtering

We had to filter some variables after building the initial predictive models to experiment further in order to identify the important variables in predicting children with high-risk of developing any IMID. More information is provided in the modelling section.

Contextual Preprocessing

The Contextual Preprocessing Guideline was followed. The details of each step are provided in continue.

Step 1. Selected Source Databases

We investigated the main purpose, data collection procedures, and reported limitations of all selected source databases at ICES and described them in section 6.1.2.2 and Appendix III. Of

course, none of the data files were perfect as they all were routinely collected real-world health administrative data. However, according to published validation studies cited in the indicated section, the sensitivity and specificity of outcome-based databases we selected were high. Overall, there was a low risk of ascertainment selection bias in our case study as the validated sensitivity and specificity in the source databases are not 100% and this risk can be tolerated as it cannot be mitigated any further.

Step 2. Criteria of Study Population

The source population of this case study was children born in Ontario from April 1, 2006, to March 31, 2012. According to Statista [127], the number of birth in Ontario in this period of time was over 838K. However, our base dataset (BORN-Niday) which formed the study population only included children born in Ontario hospitals with a gestational age of 20 weeks or greater and a birth weight of 500 grams or greater. Therefore, stillbirths and babies born outside of hospitals in Ontario were not included which were about 25K instances (2.9% of source population). Hence, our study does have a low risk of inclusion selection bias and could not be mitigated as the data relevant to the missing individuals were not accessible.

On the other hand, we had to remove about 88K invalid instances (10.5% of source population) based on the exclusion criteria. Half of the removed individuals were the ones who did not have OHIP eligibility until end of study's lookup window (March 31, 2017). In other words, we do miss the kids moving to other provinces in our study population. Therefore, there is a low risk of exclusion selection bias in our study and could not be mitigated as we only had access to Ontario's data.

The controls were not selected using matching criteria as all instances without the selected outcomes were included. So, there was no matching selection bias.

Step 3. Diagnostic Algorithms and Data Completeness

As ICES holds the most complete data records of Ontarians, we used the completest data available to look back each individual's history. However, the CIHI DAD and SDS databases only included visit records to hospitals and not to clinics; on the other hand, the OHIP database did not contain information on visits that were not claimed by the providers – the list of such claim types are listed under the OHIP data library in Appendix III (Linked Data Libraries for Case Studies). All cases were identified using a validated algorithm published in peer-reviewed articles, as they have been listed and described in section 6.1.2.3. In addition, all comorbidity variables were constructed based on expert opinion described in "Step 2. Feature Construction". However, there could be errors in entering diagnosis codes in administrative data and the validated diagnostic algorithms might also be too conservative or completely accurate. Furthermore, the type of diabetes was not specified in the ODD dataset, and we had to consider all diabetic patients under 18 years old as type 1 diabetes patients. It was still rarely possible that some diabetic patients under 18 years old were from type 2 cohort and were misdiagnosed. Therefore, there was a low risk of diagnostic bias in identifying the child cases and classifying the mothers' disease history. However, the risk of misclassification in constructing other comorbidity variables was medium due to inaccuracy of algorithms and underlying data.

Furthermore, we had to look back to inspect mother's medical history regarding the concerned outcomes. As the mother might not had been available in Ontario for all time before giving birth, we had to identify prevalent cases among mothers, instead of incident cases. For instance, the mother might had been diagnosed with the disease earlier somewhere out of Ontario, and just before birth moved to Ontario which her diagnosis claims were not complete

in the Ontario's data to match the case definition criteria. Note that the data we used contained information for decades before the earliest birth in our study population. Hence, there was a low risk of prevalence-incidence bias in the information of records included in our study dataset. On the bright side, our cases were from newborns in whom we had complete data from birth, so there was no risk of prevalence-incidence bias in selecting the cases.

Step 4. Use of Diagnostic and Procedural Codes

To construct some variables, we identified the relevant diagnostic and procedural codes and searched every child's history in the data to identify in which age they were diagnosed with the comorbidity or had the treatment. There were some concerns in using various coding standards.

One issue was with diagnosis codes that were synonym to more than one disease. For example, ICD-10 code "B97.89" was used in constructing both Rhinovirus and pneumonia (which was part of General Childhood Infections and Antibiotics) variables.

Another issue was with custom reference tables for diagnosis codes that did not fully compatible with ICD-10 codes. ICES's ICD-10 reference table that demonstrates the diagnosis codes used in DAD datasets did not exactly match the latest ICD-10 codes dictionary (icd10data.com). Some of the diagnosis codes we collected did not exist in this reference table. For instance, we asked for the ICD-10 code "A01.03" which is specifically for pneumonia; however, this code did not exist in the reference table and instead, the parent code (i.e., "A01.0") was available in the reference table which was too general and not specific to pneumonia anymore. If we did not accept the parent code, it was possible that a person diagnosed with "A01.03" was recorded as "A01.0" and then we were not including that person because the parent code contained other health conditions in addition to the special type of

pneumonia that we were concerned about. Either we picked the parent code or not, there was a risk of misclassification bias because of this issue in General Childhood Infections, Antibiotics, and Echovirus variables. In some other cases, the numbering of the diagnosis code was different between ICES's reference table and the ICD-10 dictionary. The ICD-10 codes that we could not find in the ICES' reference table or had issue with them were as following:

- A01.03 [A01.0 - not pneumonia specifically]
- A02.22 [A02.2 - no not pneumonia specifically]
- A08.39 [A08.3 - not echovirus specifically]
- A37.01 [A37.1 - not pneumonia specifically]
- A37.11 [A37.1 - not pneumonia specifically]
- A37.81 [A37.8 - not pneumonia specifically]
- A37.91 [A37.9 - not pneumonia specifically]
- A54.84 [A54.8 - not pneumonia specifically]
- B77.81 [B77.8 - not pneumonia specifically]
- B97.12 [B97.1 - not echovirus specifically]
- B97.89 -> B97.88 [same title, different numbers]
- J09.X1 [J09 - not pneumonia specifically]
- J84.11 [J84.1 - not pneumonia specifically]
- J84.2 [siblings - not pneumonia specifically]
- J84.89 [J84.8 - not pneumonia specifically]
- J95.851 [parent siblings - not pneumonia specifically]
- J95.89 -> J95.88 [same title, different numbers]
- O98.319 -> O98.3* [same title, different numbers]

- O98.32 -> O98.3* [same title, different numbers]
- O98.33 -> O98.3* [same title, different numbers]

Note that diagnostic codes in the OHIP database were a simplified 3-digit version of ICD-9 codes. Since these are not identical to the 4 to 6-digit ICD-9 codes, OHIP has its own diagnostic codes reference table.

As described in "Step 2. Feature Construction", many variables were constructed for these case studies. Most feature constructions were done based on diagnostic and procedural codes. The extracted list of codes was reviewed with domain experts multiple times to reduce the chance of mistake. However, the issues described above were still applicable, even though they might affect a very small portion of data. Hence, there was a low risk of misclassification bias in the constructed features using diagnostic and procedural codes.

Step 5. Time-Varying Variables

As some outcomes in our case study, such as IBD, had long latency period and the children might not had enough time during the follow-up window to get diagnosed, it was possible some patients developed the disease later and not included as cases in our study. However, we tried to use the latest data available to follow up the children health status to the latest possible time (they became 5 to 10 years old), it reduced the risk of missing cases. Hence, there was a medium risk of diagnostic bias in identifying patients with long latency period or spontaneous remitting or relapsing nature diseases.

All health condition variables were constructed based on the incident date of the outcome as their index date. There was no particular index date for controls and their conditions were monitored until the latest date possible in the data. Therefore, it was possible that a health condition did not get the chance to occur before outcome in cases, while it happened later for

many controls and a misleading association between the health condition and controls was detected. There was not much that we can do in this regard as the controls were not selected using the matching approach. As the follow up window in the data was not significantly long, there was a low risk of survivor treatment bias in selecting the controls and a low risk of immortal time bias in the constructed features on health conditions.

We also had location-based variables such as air quality, rurality index, and neighborhood income quintile that were constructed based on postal codes. The postal codes were not updated on a real-time basis; therefore, we observed some inconsistencies in postal code of mothers between conception and birth years. We planned to construct a version of these variables reflecting the situation during pregnancy, but because of this issue and some other concerns described in "Step 2. Feature Construction", we picked the child's postal code at birth time and looked back using gestational age to identify the related situation at conception time. This approach made the risk of misclassification bias to be low in constructing these variables. However, as we calculated the average level of particles for variables representing air quality during pregnancy and the raw data include remarkable amounts of missing values, the calculated values might not demonstrate the actual situation of air quality from conception to birth. This is in addition to the fact that the retrieved postal code at conception time might also not be very accurate. Therefore, it increases the risk of misclassification bias to high in constructing the air quality during pregnancy variables.

Step 6. Lack of Contributing Variables

Out of 128 factors we selected for this case study, only 67 factors were available as described in section 6.1.3 at the Data Selection stage. In other words, almost half of the selected factors were not available in the selected datasets obviously due to the fact that we were using

administrative data and did not include clinical data. Even though there were no factors selected related to the father of children, there was no information about them accessible in the ICES data, including their past medical issues and immigration history, as their unique identification number was not recorded. Therefore, there was a high risk of unmeasured confounding, and it cannot be mitigated as our scope was to only use administrative data and lack of accessing the information of child's father. We had already included all factors that possibly could have been selected from health administrative data since we used a very complete and rich inventory of real-world data from ICES.

In addition, some selected variables were also eliminated during the technical preprocessing – mostly among the health conditions because of having almost-unary values. These actions were necessary for the data mining methods we planned to use so that they could function and successfully build models. Considering the fact that these variables were mostly eliminated as they did not provide much information as the individuals perhaps did not have enough time to face with many of these health conditions in our data, removing them made sense. Therefore, there was a low risk of measured confounding in this study.

Step 7. Handled Missing Values

We had to truncate 0.5% of records due to high number of missing cells and imputed the rest. This is a very low number of individuals to exclude which most likely could not exclude a specific group of individuals from the analysis. Therefore, there was a trivial risk of exclusion bias in this study as a matter of truncating records with high missingness.

In addition, we decided to remove 5 variables with high amount of missingness. We did not have access in the accessible data to impute them with valid values and this action was necessary as the modelling methods we considered could not tolerate it. Even though the

number of eliminated variables were few, they represented various factors. Hence, there was a medium risk of measured confounding in this study due to removing variables with high missingness.

Step 8. Partitioned, Balanced, and Resampled Data

We included all instances available in the data so that the study population remains similar to the source population as we were using data mining techniques which were capable of handling large volume of data and did not need to take sample from the database population. Nevertheless, we had to do sampling in various points of technical preprocessing which included data partitioning, data balancing and data resampling. We used random sampling in all these attempts. However, we used software tools to implement these actions. There are always chances of imperfection in generating random numbers using built-in functions on all software platforms. We did not notice any significant distortion in the random selection occurring in RStudio. Therefore, there was a trivial risk of non-random sampling bias in this study.

As we did not use matching method in partitioning or resampling the data, there were no risk of matching bias in this study.

Lastly, we used 10-fold cross validation as the resampling method to train the predictive models. Reducing the resampling to 5-fold CV would have decreased the reliability of model's performance, and increasing to repeated 10-fold CV was not feasible due to the large number of data points and complexity of some models we used. However, as all these methods were using random sampling, there were no concern of causing selection bias in selecting the resampling method among these options.

Step 9. Model Results

We did not make any assumptions or consider any hypotheses to allow the data mining techniques discover findings in this case study. In addition, we included a variety of classifiers in the modelling stage to explore the data from different perspectives. Finally, we reported the results and findings of all experiments with reliable performance, even if the findings did not match past works in the literature. Therefore, there was a trivial risk of publication bias in this study.

Appendix VI. Multifactorial Rules in Predicting IMID Patients

Note that IMM-1 rules were extracted from the decision tree that included all input variables, IMM-4 was from the model with just the common important variables, and IMM-6 was from the model with top 20 important variables.

No.	Rule		Confidence score (model: 68%)	Applies to (out of 172K individuals)
	Conditions (Kids born in Ontario are protective or at risk of developing IMID until age of 4 to 11 years old - all factors are related to child, except mentioned explicitly)	Effect		
IMM-1-1	First antibiotics at 4 yo	Protective	82.9%	6,385
IMM-1-2	First antibiotics NOT at 4 yo (could be before, after, or never), had streptococcus pyogenes	Protective	72.2%	21,764
IMM-1-3	First antibiotics at 3 yo, no streptococcus pyogenes	Protective	73.0%	7,092
IMM-1-4	First antibiotics NOT at 3 or 4 yo (could be before, after, or never), no streptococcus pyogenes, first general childhood infection at 2 yo, birth weight equal or less than 2321 g, no respiratory infection, mom's birth month is NOT August	Risk	69.4%	49
IMM-1-5	First antibiotics NOT at 3 or 4 yo (could be before, after, or never), no streptococcus pyogenes, first general childhood infection at 2 yo, birth weight equal or less than 2321 g, had respiratory infection, baby's birth month is April, no augmentation at birth	Risk	100.0%	10
IMM-1-6	Had antibiotics but first time was NOT at 3 or 4 yo, no streptococcus pyogenes, first general childhood infection at 2 yo, birth weight greater than 2321 g	Protective	73.6%	3,011
IMM-1-7	No antibiotics, no streptococcus pyogenes, first general childhood infection at 2 yo, birth weight greater than 2321 g, had respiratory infection	Protective	66.6%	1,354
IMM-1-8	No antibiotics, no streptococcus pyogenes, first general childhood infection at 2 yo, birth weight greater than 2321 g, no respiratory infection, had gastroenteritis	Risk	70.2%	47
IMM-1-9	No streptococcus pyogenes, first general childhood infection NOT at 2 yo (could be before, after, or never), first antibiotics at 2 yo, gestational age equal or less than 32 weeks, income quintile at conception was NOT 4, mom with history of immune disease	Risk	84.6%	39

IMM-1-10	No streptococcus pyogenes, first general childhood infection NOT at 2 yo (could be before, after, or never), first antibiotics at 2 yo, gestational age greater than 32 weeks, rurality index at conception was greater than 12 (more rural)	Protective	70.2%	1,522
IMM-1-11	No streptococcus pyogenes, first general childhood infection NOT at 2 yo (could be before, after, or never), first antibiotics at 2 yo, gestational age greater than 32 weeks, rurality index at conception was equal or less than 12 (more urban), mom with no history of immune disease, baby girl	Protective	66.4%	2,944
IMM-1-12	No streptococcus pyogenes, first general childhood infection NOT at 2 yo (could be before, after, or never), first antibiotics at 2 yo, gestational age greater than 32 weeks, rurality index at conception was equal or less than 12 (more urban), mom with no history of immune disease, baby boy, no smoking during pregnancy, mom is NOT immigrant from Asia and Pacific, c-section delivery, birth month is June, PM2.5 level during pregnancy greater than 6.3 (more polluted air)	Risk	82.4%	34
IMM-1-13	No streptococcus pyogenes, first antibiotics NOT at 2 or 3 or 4 yo (could be before, after, or never), first general childhood infection NOT at 2 yo (could be before, after, or never), first respiratory infection at 2 yo, gestational age greater than 34 weeks, PM2.5 level during pregnancy equal or less than 10.6 (less polluted air), mom with no history of immune disease	Protective	67.1%	2,318
IMM-1-14	No streptococcus pyogenes, first antibiotics NOT at 2 or 3 or 4 yo (could be before, after, or never), no general childhood infection, first respiratory infection NOT at 2 yo (could be before, after, or never), first gastroenteritis at 2 yo	Protective	75.0%	116
IMM-1-15	No streptococcus pyogenes, first antibiotics NOT at 2 or 3 or 4 yo (could be before, after, or never), had general childhood infection but first time NOT at 2 yo, first respiratory infection NOT at 2 yo (could be before, after, or never), first gastroenteritis at 2 yo, mom with history of immune disease, mom is immigrant	Risk	63.9%	255
IMM-1-16	No streptococcus pyogenes, first antibiotics NOT at 2 or 3 or 4 yo (could be before, after, or never), had general childhood infection but first time NOT at 2 yo, first respiratory infection NOT at 2 yo (could be before, after, or never), first gastroenteritis at 2 yo, mom with no history of immune disease, rurality index at conception greater than 9 (more rural)	Protective	74.1%	501
IMM-1-17	No streptococcus pyogenes, had antibiotics but first time NOT at 2 or 3 or 4 yo, first general childhood infection at newborn age, had respiratory infection but first time NOT at 1 or 2 yo, first gastroenteritis NOT at 2 yo (could be before, after, or never), birth weight less than 2316 g, mom with history of immune disease	Risk	72.1%	1,459
IMM-1-18	No streptococcus pyogenes, had antibiotics but first time NOT at 3-12m or 2 or 3 or 4 yo, first general childhood infection at 1-3m of age, had respiratory infection but first time NOT at 1 or 2 yo, first	Risk	67.5%	1,018

	gastroenteritis NOT at 2 yo (could be before, after, or never), birth weight less than 2316 g, mom with history of immune disease, baby boy			
IMM-1-19	No streptococcus pyogenes, first antibiotics at 3-12m of age, first general childhood infection at 1-3m of age, had respiratory infection but first time NOT at 1 or 2 yo, first gastroenteritis NOT at 2 yo (could be before, after, or never), birth weight less than 2316 g, mom with history of immune disease	Risk	72.7%	1,425
IMM-1-20	No streptococcus pyogenes, no antibiotics, first general childhood infection at 1-3m of age, first respiratory infection NOT at 2 yo (could be before, after, or never), first gastroenteritis NOT at 2 yo (could be before, after, or never)	Risk	83.0%	7,229
IMM-1-21	No streptococcus pyogenes, no antibiotics, first general childhood infection at 3-12m of age, first respiratory infection NOT at 2 yo (could be before, after, or never), first gastroenteritis NOT at 2 yo (could be before, after, or never), mom with history of immune disease, gestational age greater than 34 weeks	Risk	81.4%	4,017
IMM-1-22	No streptococcus pyogenes, no antibiotics, no general childhood infection, first respiratory infection NOT at 2 yo (could be before, after, or never), first gastroenteritis NOT at 2 yo (could be before, after, or never), mom with history of immune disease, gestational age greater than 34 weeks	Risk	76.6%	2,743
IMM-4-1	Had streptococcus pyogenes	Protective	72.7%	22,772
IMM-4-2	No streptococcus pyogenes, first general childhood infection NOT at 2 yo (could be before, after, or never), no antibiotics, mom with history of immune disease	Risk	78.5%	11,733
IMM-4-3	No streptococcus pyogenes, first general childhood infection NOT at 2 yo (could be before, after, or never), no antibiotics, mom with no history of immune disease, baby boy	Risk	68.5%	20,625
IMM-6-1	Had streptococcus pyogenes, mom with no history of immune disease	Protective	74.9%	17,290
IMM-6-2	Had streptococcus pyogenes, mom with history of immune disease, gestational age equal or less than 32 weeks, birth weight equal or less than 2050 g	Risk	70.2%	47
IMM-6-3	Had streptococcus pyogenes, mom with history of immune disease, gestational age greater than 32 weeks, had respiratory infection, baby girl	Protective	70.1%	2,579
IMM-6-4	Had streptococcus pyogenes, mom with history of immune disease, gestational age greater than 32 weeks, had respiratory infection, baby boy, first general childhood infection NOT at 1-3m of age (could be before, after, or never)	Protective	65.1%	2,142
IMM-6-5	Had streptococcus pyogenes, mom with history of immune disease, gestational age greater than 32 weeks, had respiratory infection, baby boy, first general childhood infection at 1-3m of age, parity	Protective	62.5%	240

	greater than 0 (not first child of mom), PM2.5 level during pregnancy equal or less than 9.3 (less polluted air)			
IMM-6-6	Had streptococcus pyogenes, mom with history of immune disease, gestational age greater than 32 weeks, had respiratory infection, baby boy, first general childhood infection at 1-3m of age, parity greater than 0 (not first child of mom), PM2.5 level during pregnancy greater than 9.3 (more polluted air)	Risk	72.1%	43
IMM-6-7	Had streptococcus pyogenes, mom with history of immune disease, gestational age greater than 32 weeks, had respiratory infection, baby boy, first general childhood infection at 1-3m of age, parity is 0 (first child of mom), no gastroenteritis, NO2 level during pregnancy less than 11.5 (less polluted air)	Protective	68.2%	44
IMM-6-8	Had streptococcus pyogenes, mom with history of immune disease, gestational age greater than 32 weeks, had respiratory infection, baby boy, first general childhood infection at 1-3m of age, parity is 0 (first child of mom), no gastroenteritis, NO2 level during pregnancy equal or greater than 11.5 (more polluted air)	Risk	85.7%	21
IMM-6-9	Had streptococcus pyogenes, mom with history of immune disease, gestational age greater than 32 weeks, no respiratory infection, first general childhood infection at 3-12m of age	Risk	64.2%	53
IMM-6-10	Had streptococcus pyogenes, mom with history of immune disease, gestational age greater than 32 weeks, no respiratory infection, first general childhood infection at newborn age	Risk	85.7%	7
IMM-6-11	Had streptococcus pyogenes, mom with history of immune disease, gestational age greater than 32 weeks, no respiratory infection, first general childhood infection NOT at newborn or 1-3m or 3-12m or 2 yo (could be before, between, after, or never)	Protective	72.6%	146
IMM-6-12	Had streptococcus pyogenes, mom with history of immune disease, gestational age greater than 32 weeks, no respiratory infection, first general childhood infection at 2 yo, parity equal or less than 1 (first or second child of mom), had antibiotics	Risk	68.8%	16
IMM-6-13	No streptococcus pyogenes, first general childhood infection at 2 yo, gestational age equal or less than 36, rurality index at conception greater than 6 (more rural), PM2.5 level during pregnancy equal or less than 8.2 (less polluted air)	Protective	75.0%	112
IMM-6-14	No streptococcus pyogenes, first general childhood infection at 2 yo, gestational age equal or less than 36, rurality index at conception equal or less than 6 (more urban), no respiratory infection	Risk	61.0%	77
IMM-6-15	No streptococcus pyogenes, first general childhood infection at 2 yo, gestational age greater than 36, had antibiotics	Protective	74.1%	3,767

IMM-6-16	No streptococcus pyogenes, first general childhood infection at 2 yo, gestational age greater than 36, no antibiotics, first respiratory infection at 2 yo	Protective	66.8%	1,250
IMM-6-17	No streptococcus pyogenes, first general childhood infection at 2 yo, gestational age greater than 36, no antibiotics, had respiratory infection but first time NOT at 2 yo, birth weight greater than 3157 g	Protective	86.7%	60
IMM-6-18	No streptococcus pyogenes, first general childhood infection at 2 yo, gestational age greater than 36, no antibiotics, had respiratory infection but first time NOT at 2 yo, birth weight equal or less than 3157 g, PM2.5 level during pregnancy greater than 4.6 (more polluted air)	Risk	90.9%	11
IMM-6-19	No streptococcus pyogenes, first general childhood infection at 1 yo, had antibiotics, had respiratory infection but first time NOT at 1 or 2 yo	Protective	82.8%	1,518
IMM-6-20	No streptococcus pyogenes, first general childhood infection at 1 yo, had antibiotics, first respiratory infection at 2 yo, mom with no history of immune disease, birth weight greater than 2058 g	Protective	69.2%	1,264
IMM-6-21	No streptococcus pyogenes, first general childhood infection at 1 yo, had antibiotics, first respiratory infection at 2 yo, mom with no history of immune disease, birth weight equal or less than 2058 g, parity is 0 (first child of mom), NO2 level during pregnancy greater than 6.6818 (more polluted air)	Risk	82.6%	23
IMM-6-22	No streptococcus pyogenes, first general childhood infection at 1 yo, had antibiotics, no respiratory infection, gestational age equal or less than 32	Risk	87.3%	71
IMM-6-23	No streptococcus pyogenes, first general childhood infection at 1 yo, had antibiotics, no respiratory infection, gestational age greater than 32, PM2.5 level during pregnancy equal or less than 3.4 (less polluted air)	Protective	65.1%	502
IMM-6-24	No streptococcus pyogenes, first general childhood infection at 1 yo, had antibiotics, no respiratory infection, gestational age greater than 32, PM2.5 level during pregnancy greater than 3.4 (more polluted air), mom with history of immune disease, parity equal or less than 2 (first three child of mom), baby boy, no gastroenteritis	Risk	69.1%	178
IMM-6-25	No streptococcus pyogenes, first general childhood infection at 1 yo, had antibiotics, no respiratory infection, gestational age greater than 32, PM2.5 level during pregnancy greater than 3.4 (more polluted air), mom with history of immune disease, parity equal or less than 2 (first three child of mom), baby boy, had gastroenteritis, birth weight greater than 2720 g	Risk	60.3%	73
IMM-6-26	No streptococcus pyogenes, first general childhood infection NOT at 1 or 2 yo, had antibiotics, mom with no history of immune disease, birth weight equal or less than 1829 g	Risk	70.4%	1,114

Appendix VII. Multifactorial Rules in Predicting Asthmatic Patients

Note that ASTH-1 rules were extracted from the decision tree that included all input variables.

No.	Rule		Confidence score (model: 69%)	Applies to (out of 167K individuals)
	Conditions (Kids born in Ontario are protective or at risk of developing Asthma until age of 4 to 11 years old - all factors are related to child, except mentioned explicitly)	Effect		
ASTH-1-1	First antibiotics at 4 yo	Protective	82.2%	6,006
ASTH-1-2	First antibiotics NOT at 4 yo (could be before, after, or never), had streptococcus pyogenes, mom with no history of asthma	Protective	74.1%	17,314
ASTH-1-3	First antibiotics NOT at 4 yo (could be before, after, or never), had streptococcus pyogenes, mom has history of asthma, first respiratory infection NOT at 2 yo (could be before, after, or never), first general childhood infection at 1 yo	Protective	72.4%	500
ASTH-1-4	First antibiotics NOT at 4 yo (could be before, after, or never), had streptococcus pyogenes, mom has history of asthma, first respiratory infection NOT at 2 yo (could be before, after, or never), first general childhood infection NOT at 1 yo (could be before, after, or never), birth weight equal or less than 2439 g, income quintile during pregnancy greater than 1	Risk	82.5%	40
ASTH-1-5	First antibiotics at 3 yo, no streptococcus pyogenes	Protective	73.4%	6,883
ASTH-1-6	First antibiotics NOT at 3 or 4 yo (could be before, after, or never), no streptococcus pyogenes, first general childhood infection at 2 yo, birth weight greater than 2989 g	Protective	70.8%	4,365
ASTH-1-7	First antibiotics NOT at 3 or 4 yo (could be before, after, or never), no streptococcus pyogenes, first general childhood infection at 2 yo, birth weight equal or less than 2989 g, birth month is NOT Dec, mom is immigrant	Risk	69.1%	55
ASTH-1-8	First antibiotics at 2 yo, no streptococcus pyogenes, first general childhood infection NOT at 2 yo (could be before, after, or never), gestational age greater than 34 weeks, mom with no history of asthma, rurality index during pregnancy greater than 11 (more rural), first respiratory infection NOT at 1-3m of age (could be before, after, or never)	Protective	75.0%	1,048

ASTH-1-9	First antibiotics at 2 yo, no streptococcus pyogenes, first general childhood infection NOT at 2 yo (could be before, after, or never), gestational age greater than 34 weeks, mom with no history of asthma, rurality index during pregnancy greater than 11 (more rural), first respiratory infection at 1-3m of age, birth month is NOT Feb or Sep, had rhinitis	Risk	76.9%	13
ASTH-1-10	First antibiotics at 2 yo, no streptococcus pyogenes, first general childhood infection NOT at 2 yo (could be before, after, or never), gestational age greater than 34 weeks, mom with no history of asthma, rurality index during pregnancy greater than 11 (more rural), first respiratory infection at 1-3m of age, birth month is NOT Feb or Sep, no rhinitis, income quintile at birth NOT 3, no labour during birth	Risk	78.6%	14
ASTH-1-11	First antibiotics at 2 yo, no streptococcus pyogenes, first general childhood infection NOT at 2 yo (could be before, after, or never), gestational age greater than 34 weeks, mom with no history of asthma, rurality index during pregnancy equal or less than 11 (more urban), mom smoking during pregnancy	Protective	71.5%	431
ASTH-1-12	First antibiotics at 2 yo, no streptococcus pyogenes, gestational age greater than 34 weeks, mom with no history of asthma, rurality index during pregnancy equal or less than 11 (more urban), mom NOT smoking during pregnancy, first gastroenteritis NOT at 3-12m of age (could be before, after, or never), mom is immigrant from Asia and Pacific, first general childhood infection at newborn age	Risk	61.3%	75
ASTH-1-13	First antibiotics at 2 yo, no streptococcus pyogenes, first general childhood infection NOT at 2 yo (could be before, after, or never), gestational age greater than 34 weeks, mom with no history of asthma, rurality index during pregnancy equal or less than 11 (more urban), mom NOT smoking during pregnancy, first gastroenteritis NOT at 3-12m of age (could be before, after, or never), mom is NOT immigrant from Asia and Pacific, mom with no history of any immune disease, baby girl, first respiratory infection NOT at 2 yo (could be before, after, or never), mom is immigrant, mom's birth month is NOT Mar	Protective	63.5%	392
ASTH-1-14	First antibiotics at 2 yo, no streptococcus pyogenes, first general childhood infection NOT at 2 yo (could be before, after, or never), gestational age greater than 34 weeks, mom with no history of asthma, rurality index during pregnancy equal or less than 11 (more urban), mom NOT smoking during pregnancy, first gastroenteritis NOT at 3-12m of age (could be before, after, or never), mom is NOT immigrant from Asia and Pacific, mom with no history of any immune disease, baby girl, first respiratory infection NOT at 2 yo (could be before, after, or never), mom is immigrant, mom's birth month is Mar	Risk	66.7%	30
ASTH-1-15	First antibiotics at 2 yo, no streptococcus pyogenes, first general childhood infection NOT at 2 yo (could be before, after, or never), gestational age greater than 34 weeks, mom with no history of	Protective	73.0%	1,136

	asthma, rurality index during pregnancy equal or less than 11 (more urban), mom NOT smoking during pregnancy, first gastroenteritis NOT at 3-12m of age or 2 yo (maybe before, after, between, never), mom is NOT immigrant from Asia and Pacific, mom with no history of any immune disease, baby girl, first respiratory infection NOT at 2 yo (could be before, after, or never), mom is NOT immigrant			
ASTH-1-16	First antibiotics at 2 yo, no streptococcus pyogenes, first general childhood infection NOT at 2 yo (could be before, after, or never), gestational age greater than 34 weeks, mom with no history of asthma, rurality index during pregnancy equal or less than 11 (more urban), mom NOT smoking during pregnancy, mom is NOT immigrant from Asia and Pacific, mom with no history of any immune disease, baby girl, first respiratory infection NOT at 2 yo (could be before, after, or never), mom is NOT immigrant, first gastroenteritis at 2 yo, no newborn resuscitation	Protective	68.0%	100
ASTH-1-17	First antibiotics at 2 yo, no streptococcus pyogenes, first general childhood infection NOT at 2 yo (could be before, after, or never), gestational age greater than 34 weeks, mom with no history of asthma, rurality index during pregnancy equal or less than 11 (more urban), mom NOT smoking during pregnancy, mom is NOT immigrant from Asia and Pacific, mom with no history of any immune disease, baby girl, first respiratory infection NOT at 2 yo (could be before, after, or never), mom is NOT immigrant, first gastroenteritis at 2 yo, had newborn resuscitation	Risk	66.7%	30
ASTH-1-18	First antibiotics at 2 yo, no streptococcus pyogenes, first general childhood infection NOT at 2 yo (could be before, after, or never), gestational age greater than 34 weeks, mom with no history of asthma, rurality index during pregnancy equal or less than 11 (more urban), mom NOT smoking during pregnancy, first gastroenteritis at 3-12m of age, PM2.5 level during pregnancy greater than 10 (more polluted air)	Risk	64.5%	152
ASTH-1-19	Had antibiotics but first time NOT at 2 or 3 or 4 yo, no streptococcus pyogenes, first general childhood infection NOT at 2 yo (could be before, after, or never), first respiratory infection at 2 yo, mom is immigrant from Asia and Pacific, mom with history of immune disease	Risk	71.8%	39
ASTH-1-20	Had antibiotics but first time NOT at 2 or 3 or 4 yo, no streptococcus pyogenes, first general childhood infection NOT at 2 yo (could be before, after, or never), first respiratory infection at 2 yo, mom is NOT immigrant from Asia and Pacific, PM2.5 level during pregnancy equal or less than 8.9 (less polluted air)	Protective	70.6%	1,861
ASTH-1-21	No antibiotics, no streptococcus pyogenes, first general childhood infection NOT at 2 yo (could be before, after, or never), first respiratory infection at 2 yo, mom is immigrant, mom with no history of immune disease, NO2 level during pregnancy greater than 7.2523 (more polluted air)	Risk	68.2%	66
ASTH-1-22	First antibiotics NOT at 2 or 3 or 4 yo (could be before, after, or never), no streptococcus pyogenes, first general childhood infection NOT at 2 yo (could be before, after, or never), first respiratory	Risk	66.2%	148

	infection NOT at 2 yo (could be before, after, or never), first gastroenteritis at 2 yo, mom with history of asthma, mom is immigrant			
ASTH-1-23	First antibiotics NOT at 2 or 3 or 4 yo (could be before, after, or never), no streptococcus pyogenes, first general childhood infection NOT at 2 yo (could be before, after, or never), first respiratory infection NOT at 2 yo (could be before, after, or never), first gastroenteritis at 2 yo, mom with history of asthma, mom is NOT immigrant, baby's birth month is Sep	Risk	66.7%	54
ASTH-1-24	First antibiotics NOT at 2 or 3 or 4 yo (could be before, after, or never), no streptococcus pyogenes, first general childhood infection NOT at 2 yo (could be before, after, or never), first gastroenteritis at 2 yo, mom with no history of asthma, no respiratory infection	Protective	77.9%	258
ASTH-1-25	First antibiotics NOT at 2 or 3 or 4 yo (could be before, after, or never), no streptococcus pyogenes, first general childhood infection NOT at 2 yo (could be before, after, or never), had respiratory infection but first time NOT at 2 yo, first gastroenteritis at 2 yo, mom with no history of asthma, rurality index during pregnancy greater than 7 (more rural)	Protective	74.2%	596
ASTH-1-26	First antibiotics NOT at 2 or 3 or 4 yo (could be before, after, or never), no streptococcus pyogenes, first general childhood infection NOT at 2 yo (could be before, after, or never), first gastroenteritis at 2 yo, mom with no history of asthma, rurality index during pregnancy equal or less than 7 (more urban), first respiratory infection at 1-3m of age, had newborn resuscitation	Risk	65.0%	80
ASTH-1-27	First antibiotics NOT at 2 or 3 or 4 yo (could be before, after, or never), no streptococcus pyogenes, first general childhood infection NOT at 2 yo (could be before, after, or never), first gastroenteritis at 2 yo, mom with no history of asthma, rurality index during pregnancy equal or less than 7 (more urban), had respiratory infection but first time NOT at 1-3m or 2 yo, baby girl	Protective	66.2%	1,091
ASTH-1-28	First antibiotics NOT at 2 or 3 or 4 yo (could be before, after, or never), no streptococcus pyogenes, first general childhood infection at 1 yo, no respiratory infection, first gastroenteritis NOT at 2 yo (could be before, after, or never), gestational age equal or less than 36 weeks	Risk	72.1%	362
ASTH-1-29	First antibiotics NOT at 2 or 3 or 4 yo (could be before, after, or never), no streptococcus pyogenes, first general childhood infection at 1 yo, had respiratory infection but first time NOT at 1 or 2 yo, first gastroenteritis NOT at 2 yo (could be before, after, or never), gestational age equal or less than 36 weeks	Protective	76.7%	103
ASTH-1-30	First antibiotics NOT at 2 or 3 or 4 yo (could be before, after, or never), no streptococcus pyogenes, first general childhood infection at 1 yo, first respiratory infection at 1 yo, first gastroenteritis NOT at 2 yo (could be before, after, or never), gestational age equal or less than 32 weeks	Risk	72.7%	209

ASTH-1-31	Had antibiotics but first time NOT at 2 or 3 or 4 yo, no streptococcus pyogenes, first general childhood infection at 1 yo, first respiratory infection NOT at 2 yo (could be before, after, or never), first gastroenteritis NOT at 2 yo (could be before, after, or never), gestational age greater than 36 weeks, rurality index during pregnancy greater than 6 (more rural), mom with no history of asthma, baby girl, mom's birth month is NOT Aug	Protective	75.3%	1,068
ASTH-1-32	First antibiotics NOT at 2 or 3 or 4 yo (could be before, after, or never), no streptococcus pyogenes, first general childhood infection at 1 yo, first respiratory infection NOT at 2 yo (could be before, after, or never), first gastroenteritis NOT at 2 yo (could be before, after, or never), gestational age greater than 36 weeks, rurality index during pregnancy greater than 6 (more rural), mom with history of asthma, baby girl, mom's birth month is Aug	Risk	64.3%	28
ASTH-1-33	No antibiotics, no streptococcus pyogenes, first general childhood infection at 1 yo, no respiratory infection, first gastroenteritis NOT at 2 yo (could be before, after, or never), gestational age greater than 36 weeks, rurality index during pregnancy equal or less than 6 (more urban)	Risk	76.5%	439
ASTH-1-34	Had antibiotics but first time NOT at 2 or 3 or 4 yo, no streptococcus pyogenes, first general childhood infection at 1 yo, no respiratory infection, first gastroenteritis NOT at 2 yo (could be before, after, or never), gestational age greater than 36 weeks, rurality index during pregnancy is 6	Risk	68.9%	106
ASTH-1-35	First antibiotics NOT at 2 or 3 or 4 yo (could be before, after, or never), no streptococcus pyogenes, first general childhood infection at 1 yo, had respiratory infection but first time NOT at 1 or 2 yo, first gastroenteritis NOT at 2 yo (could be before, after, or never), gestational age greater than 36 weeks, rurality index during pregnancy equal or less than 6 (more urban)	Protective	82.6%	811
ASTH-1-36	Had antibiotics but first time NOT at 3-12m or 1 or 2 or 3 or 4 yo, no streptococcus pyogenes, first general childhood infection at 1 yo, first respiratory infection at 1 yo, first gastroenteritis NOT at 2 yo (could be before, after, or never), gestational age greater than 36 weeks, rurality index during pregnancy equal or less than 6 (more urban), mom with history of asthma	Protective	85.5%	69
ASTH-1-37	First antibiotics at 3-12m of age, no streptococcus pyogenes, first general childhood infection at 1 yo, first respiratory infection at 1 yo, first gastroenteritis NOT at 2 yo (could be before, after, or never), gestational age greater than 36 weeks, rurality index during pregnancy equal or less than 6 (more urban), mom with history of asthma	Risk	80.0%	10
ASTH-1-38	First antibiotics NOT at 2 or 3 or 4 (could be before, after, or never), no streptococcus pyogenes, first general childhood infection at 1 yo, first respiratory infection at 1 yo, first gastroenteritis NOT at 2 yo (could be before, after, or never), gestational age greater than 36 weeks, rurality index during pregnancy equal or less than 6 (more urban), mom with no history of asthma, parity greater than 0	Protective	77.0%	100

	(not the first child of mom), mom had intention to breastfeed, NO2 level during pregnancy equal or less than 18.3788 (less polluted air), baby boy, mom's age at birth equal or less than 25, PM2.5 level during pregnancy equal or less than 10.3 (less polluted air)			
ASTH-1-39	First antibiotics NOT at 2 or 3 or 4 (could be before, after, or never), no streptococcus pyogenes, first general childhood infection at 1 yo, first respiratory infection at 1 yo, first gastroenteritis NOT at 2 yo (could be before, after, or never), gestational age greater than 36 weeks, rurality index during pregnancy equal or less than 6 (more urban), mom with no history of asthma, parity greater than 0 (not the first child of mom), mom had intention to breastfeed, NO2 level during pregnancy equal or less than 18.3788 (less polluted air), baby boy, mom's age at birth equal or less than 25, PM2.5 level during pregnancy greater than 10.3 (more polluted air)	Risk	88.9%	9
ASTH-1-40	First antibiotics NOT at 2 or 3 or 4 (could be before, after, or never), no streptococcus pyogenes, first general childhood infection at 1 yo, first respiratory infection at 1 yo, first gastroenteritis NOT at 2 yo (could be before, after, or never), gestational age greater than 36 weeks, rurality index during pregnancy equal or less than 6 (more urban), mom with no history of asthma, parity greater than 0 (not the first child of mom), mom had intention to breastfeed, NO2 level during pregnancy equal or less than 18.3788 (less polluted air), baby girl, baby's birth month is Jun	Protective	75.5%	102
ASTH-1-41	First antibiotics at 3-12m of age, no streptococcus pyogenes, first general childhood infection at 1 yo, first respiratory infection at 1 yo, first gastroenteritis NOT at 2 yo (could be before, after, or never), gestational age greater than 36 weeks, rurality index during pregnancy equal or less than 6 (more urban), mom with no history of asthma, parity is 0 (first child of mom), mom is NOT immigrant	Risk	71.4%	28
ASTH-1-42	No antibiotics, no streptococcus pyogenes, first general childhood infection at 1 yo, first respiratory infection at 1 yo, had gastroenteritis but first time is NOT at 2 yo, gestational age greater than 36 weeks, rurality index during pregnancy equal or less than 6 (more urban), mom with no history of asthma, parity is 0 (first child of mom), mom is NOT immigrant, baby girl, mom's birth month is NOT Jul, income quintile at birth is 3	Risk	85.0%	20
ASTH-1-43	No antibiotics, no streptococcus pyogenes, first general childhood infection at 1 yo, first respiratory infection at 1 yo, no gastroenteritis, gestational age greater than 36 weeks, rurality index during pregnancy equal or less than 6 (more urban), mom with no history of asthma, parity is 0 (first child of mom), mom is NOT immigrant, baby girl, mom's birth month is NOT Jul, income quintile at birth is 3, income quintile during pregnancy is NOT 2, baby's birth month is NOT May, birth was assisted (forceps or vacuum)	Protective	85.7%	7
ASTH-1-44	No antibiotics, no streptococcus pyogenes, first general childhood infection at 1 yo, first respiratory infection at 1 yo, no gastroenteritis, gestational age greater than 36 weeks, rurality index during	Risk	71.4%	35

	pregnancy equal or less than 6 (more urban), mom with no history of asthma, parity is 0 (first child of mom), mom is NOT immigrant, baby girl, mom's birth month is NOT Jul, income quintile at birth is 3, income quintile during pregnancy is NOT 2, baby's birth month is NOT May, birth was NOT assisted (with forceps or vacuum)			
ASTH-1-45	No antibiotics, no streptococcus pyogenes, first general childhood infection at 1-3m of age, first respiratory infection NOT at 2 yo, first gastroenteritis NOT at 2 yo, mom with history of asthma	Risk	89.1%	1,574
ASTH-1-46	No antibiotics, no streptococcus pyogenes, first general childhood infection at newborn age, first respiratory infection NOT at 2 yo, first gastroenteritis NOT at 2 yo, mom with history of asthma	Risk	90.6%	566
ASTH-1-47	No antibiotics, no streptococcus pyogenes, first general childhood infection at 3-12m of age, first respiratory infection NOT at 2 yo, first gastroenteritis NOT at 2 yo, mom with history of asthma	Risk	84.9%	3,235
ASTH-1-48	No antibiotics, no streptococcus pyogenes, had general childhood infection but first time at or after 2 yo, first respiratory infection NOT at 2 yo, first gastroenteritis NOT at 2 yo, mom with history of asthma	Protective	75.9%	191
ASTH-1-49	No antibiotics, no streptococcus pyogenes, no general childhood infection, first respiratory infection NOT at 2 yo, first gastroenteritis NOT at 2 yo, mom with history of asthma, baby's birth month is NOT Feb	Risk	80.0%	1,986
ASTH-1-50	No antibiotics, no streptococcus pyogenes, no general childhood infection, first respiratory infection NOT at 2 yo, first gastroenteritis NOT at 2 yo, mom with history of asthma, baby's birth month is Feb, NO2 level at birth greater than 6.7238 (more polluted air)	Risk	71.2%	125
ASTH-1-51	First antibiotics at 1 yo, no streptococcus pyogenes, first general childhood infection NOT at 1 or 2 yo, no respiratory infection, first gastroenteritis NOT at 2 yo, mom with history of asthma, mom intended to breastfeed	Risk	76.1%	71
ASTH-1-52	First antibiotics at 1 yo, no streptococcus pyogenes, first general childhood infection NOT at 1 or 2 yo, no respiratory infection, first gastroenteritis NOT at 2 yo, mom with history of asthma, mom intended NOT to breastfeed	Protective	72.7%	11
ASTH-1-53	Had antibiotics but first time NOT at 2 or 3 or 4 yo, no streptococcus pyogenes, first general childhood infection NOT at 1 or 2 yo, had respiratory infection but first time NOT at 2 yo, first gastroenteritis NOT at 2 yo, mom with history of asthma, mom is immigrant from Asia and Pacific	Risk	76.8%	786
ASTH-1-54	Had antibiotics but first time NOT at 2 or 3 or 4 yo, no streptococcus pyogenes, first general childhood infection NOT at 1 or 2 yo, had respiratory infection but first time NOT at 2 yo, first gastroenteritis NOT at 2 yo, mom with history of asthma, mom is NOT immigrant from Asia and Pacific, birth weight equal or less than 2497 g	Risk	81.1%	249

ASTH-1-55	No antibiotics, no streptococcus pyogenes, first general childhood infection NOT at 1 or 2 yo, first respiratory infection NOT at 2 yo, first gastroenteritis NOT at 2 yo, mom with no history of asthma, rurality index during pregnancy greater than 10 (more rural), PM2.5 level during pregnancy equal or less than 2.3 (less polluted air)	**Protective**	97.1%	68

Appendix VIII. Multifactorial Rules in Predicting IBD Patients

Note that IBD-U rules were extracted from the decision tree built on the undersampled data that included all input variables.

No.	Rule		Confidence score (model: 66%)	Applies to (out of 234 individuals)
	Conditions (Kids born in Ontario are protective or at risk of developing IBD until age of 4 to 11 years old - all factors are related to child, except mentioned explicitly)	Effect		
IBD-U-1	No gastroenteritis, born in July	Protective	100%	7
IBD-U-2	No gastroenteritis, not born in July, first general childhood infection at 1 yo	Risk	66.70%	12
IBD-U-3	Had gastroenteritis, born in March, gestational age equal or less than 39 weeks	Risk	75%	8
IBD-U-4	Had gastroenteritis, born in March, gestational age greater than 39 weeks	Protective	100%	6
IBD-U-5	Had gastroenteritis, not born in March and December, mom had a maternal health problem	Risk	94.40%	18
IBD-U-6	Had gastroenteritis, not born in March and December, mom had no labour at birth	Risk	81.80%	11
IBD-U-7	Had gastroenteritis, not born in March and December, mom had labour at birth, first general childhood infection NOT at 1 yo (could be before, after, or never)	Risk	70.80%	72
IBD-U-8	Had gastroenteritis, not born in March and December, mom had labour at birth, first general childhood infection at 1 yo, PM2.5 level at birth was equal or less than 8.9 (less polluted air)	Protective	100%	7
IBD-U-9	Had gastroenteritis, not born in March and December, mom had labour at birth, first general childhood infection at 1 yo, PM2.5 level at birth was greater than 8.9 (more polluted air)	Risk	83.30%	6

Appendix IX. Multifactorial Rules in Predicting SARD Patients

Note that SARD-1 rules were extracted from the decision tree that included all input variables.

No.	Rule		Confidence score (model: 66%)	Applies to (out of 3.4K individuals)
	Conditions (Kids born in Ontario are protective or at risk of developing SARDs until age of 4 to 11 years old - all factors are related to child, except mentioned explicitly)	Effect		
SARD-1-1	Mother with history of SARD	Risk	79.9%	443
SARD-1-2	Mother with no history of SARD, had antibiotics	Protective	64.0%	2,037
SARD-1-3	Mother with no history of SARD, no antibiotics, had streptococcus pyogenes	Protective	82.9%	41
SARD-1-4	Mother with no history of SARD, no antibiotics, no streptococcus pyogenes, first respiratory infection NOT at 2 yo (could be before, after, or never)	Risk	69.8%	870
SARD-1-5	Mother with no history of SARD, no antibiotics, no streptococcus pyogenes, first respiratory infection at 2 yo	Protective	71.0%	31